DOI: 10.5748/18CONTECSI/PSE/GOV/6853

CONNECTED OPEN GOVERNMENT DATA IN BIG DATA: PROPOSAL OF A CONCEPTUAL FRAMEWORK

Leonardo Ferreira Da Silva - INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS - Orcid: https://orcid.org/0000-0003-2787-9660

Eduardo Amadeu Dutra Moresi - UNIVERSIDADE CATÓLICA DE BRASÍLIA - Orcid: https://orcid.org/0000-0001-6058-3883

This work addresses the communication and interoperability between government environments that can improve the use of available information. Government Open Data can offer a lot of information, being useful in enriching knowledge as long as used in a correlated way. Especially in the Education's area of the federal government, most of the open data is available in a dispersed form over the internet. Therefore, this work presents how to develop a conceptual framework for Government Open Data Connected in Big Data. Deductive reasoning was performed to have sufficient subsidies for proposing a preliminary conceptual framework for the acquisition and use of open government data of interest in the educational area, later on inductive reasoning was used to verify the framework. The survey was conducted with experts in open educational data dissemination. As main results it was possible to identify the environment, the structure, the elements and the variables that involve the collection and dissemination of data that compose the basic structure of an environment that receives and consume structured and unstructured data. It is hoped that the proposed conceptual framework will become a reference point for future research in open government data and new technologies such as big data and data lake and will allow not only the educational field, but also other spheres of government to implement it. The main contribution is the conceptual framework for collecting, preparing and disseminating data about Education subject area.

Keywords: Open data, Open government data, Big Data, Data Lake, Conceptual framework, Education

DADOS ABERTOS GOVERNAMENTAIS CONECTADOS EM BIG DATA: PROPOSTA DE UM FRAMEWORK CONCEITUAL

Este trabalho aborda a importância da comunicação e da interoperabilidade entre ambientes governamentais que favorecem o melhor uso da informação disponível. Os Dados Abertos Governamentais podem oferecer uma infinidade de informações sendo úteis no enriquecimento do conhecimento desde que usados de forma relacionada. Em especial na área da Educação do governo federal, a maioria dos dados abertos estão disponibilizados de forma dispersa pela internet. Neste sentido, este trabalho apresenta como desenvolver um framework conceitual para Dados Abertos Governamentais Conectados em Big Data. Realizou-se um raciocínio dedutivo para ter subsídios para a proposição de um framework conceitual para aquisição e utilização de dados abertos governamentais de interesse da área educacional, posteriormente foi utilizado o raciocínio indutivo para verificação do framework A pesquisa foi realizada junto a especialistas em disseminação de dados abertos educacionais. Como principais resultados foi possível identificar o ambiente, a estrutura, os elementos e as variáveis que envolvem a coleta e a disseminação de dados que compõem a estrutura básica de um ambiente que recepcione dados estruturados e não estruturados para serem consumidos. Espera-se que o framework conceitual proposto venha a ser um ponto de referência para pesquisas futuras em dados abertos governamentais e novas tecnologias como big data e data lake e permita que, não somente, a área educacional, mas também outras esferas do governo possam implementá-lo. A principal contribuição é o framework conceitual para a coleta, preparação e divulgação de dados da área de Educação.

Palavras-chave: Dados abertos, Dados abertos governamentais, Big data, Data Lake, Framework conceitual, Educação

Dados abertos governamentais conectados em big data: proposta de um *framework* conceitual

Leonardo Ferreira da Silva – ORCID (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – leo.lfs27@gmail.com)
Eduardo Amadeu Dutra Moresi – ORCID 0000-0001-6058-3883 (Universidade Católica de Brasília – moresi@p.ucb.br)

Resumo

Objetivo do estudo

Este trabalho aborda a importância da comunicação e da interoperabilidade entre ambientes governamentais que favorecem o melhor uso da informação disponível. Os Dados Abertos Governamentais podem oferecer uma infinidade de informações sendo úteis no enriquecimento do conhecimento desde que usados de forma relacionada.

Relevância/originalidade

Em especial na área da Educação do governo federal, a maioria dos dados abertos estão disponibilizados de forma dispersa pela internet. Neste sentido, este trabalho apresenta como desenvolver um *framework* conceitual para Dados Abertos Governamentais Conectados em *Big Data*.

Metodologia/abordagem

Realizou-se um raciocínio dedutivo para ter subsídios para a proposição de um framework conceitual para aquisição e utilização de dados abertos governamentais de interesse da área educacional, posteriormente foi utilizado o raciocínio indutivo para verificação do framework. A pesquisa foi realizada junto a especialistas em disseminação de dados abertos educacionais.

Principais resultados

Como principais resultados foi possível identificar o ambiente, a estrutura, os elementos e as variáveis que envolvem a coleta e a disseminação de dados que compõem a estrutura básica de um ambiente que recepcione dados estruturados e não estruturados para serem consumidos.

Contribuições teóricas/metodológicas

Espera-se que o *framework* conceitual proposto venha a ser um ponto de referência para pesquisas futuras em dados abertos governamentais e novas tecnologias como *big data* e *data lake* e permita que, não somente, a área educacional, mas também outras esferas do governo possam implementá-lo.

Contribuições sociais/para a gestão

A principal contribuição é o framework conceitual para a coleta, preparação e divulgação de dados da área de Educação.

Palavras-chave: Dados abertos. Dados abertos governamentais. *Big data*. *Data Lake*. *Framework* conceitual. Educação.

Connected open government data in big data: proposal of a conceptual framework

Abstract

Study goals

This work addresses the communication and interoperability between government environments that can improve the use of available information. Government Open Data can offer a lot of information, being useful in enriching knowledge as long as used in a correlated way.

Relevance / originality
Especially in the Education's area of the federal government, most of the open data is available in a dispersed form over the internet. Therefore, this work presents how to develop a conceptual framework for Government Open Data Connected in Big Data.

Methodology / approach
Deductive reasoning was performed to have sufficient subsidies for proposing a preliminary conceptual framework for the acquisition and use of open government data of interest in the educational area, later on inductive reasoning was used to verify the framework. The survey was conducted with experts in open educational data dissemination.

Main results
As main results it was possible to identify the environment, the structure, the elements and the variables that involve the collection and dissemination of data that compose the basic structure of an environment that receives and consume structured and unstructured data.

Theoretical / methodological contributions
It is hoped that the proposed conceptual framework will become a reference point for future research in open government data and new technologies such as big data and data lake and will allow not only the educational field, but also other spheres of government to implement it.

Social / management contributions
The main contribution is the conceptual framework for collecting, preparing and disseminating data about Education subject area.

Keywords: Open data. Open government data. Big Data. Data Lake. Conceptual framework. Education.

1. Introdução

As barreiras na comunicação podem significar o retardamento tecnológico da sociedade. Oliveira (2016), ao revisitar o mito sobre a Torre de Babel descreve que enquanto a comunicação era comum entre os povos, possibilitou a construção da torre, porém, posteriormente, dificultada pela ira divina na criação de diferentes línguas entre os povos, essa lenda denota a importância da comunicação, principalmente em larga escala. Com a barreira da língua, os babilônicos desistiram do projeto da torre.

Nos últimos anos, a revolução digital criou uma explosão sem precedentes na capacidade de adquirir, armazenar, manipular e transmitir instantaneamente vastos e complexos volumes de dados (Science International, 2015). Casaes (2019) define dados abertos como aqueles com acesso livre não importando a plataforma em que estejam disponíveis. Attard et al (2015) asseveram que dados abertos não devem possuir restrição de uso. Nesse contexto, acrescenta-

se aspectos legais e técnicos que devem ser seguidos para que essa disponibilização seja realizada.

O uso de dados abertos, particularmente os produzidos pelo próprio governo (dados abertos governamentais), é a porta de entrada para fornecer canais de comunicação com a sociedade civil, servindo também para municiar a comunidade cientifica em suas pesquisas bem como auxiliar os gestores públicos na melhoria e desenvolvimento de programas governamentais. Entenda-se conjuntos de dados como uma combinação de um ou mais arquivos que contenham algum tipo de informação, esses são criados e disponibilizados sem que exista a preocupação de haver uma interoperabilidade entre eles.

O cenário dos dados abertos educacionais ainda é pouco explorado. Santos, Ferreira e Miranda (2017) constataram que artigos publicados com base em dados abertos voltados para área da educação ainda estão se consolidando. Werle (2014) destaca o uso de tecnologias digitais com intuito de melhorar o gerenciamento do sistema de ensino, assim, beneficiando as políticas públicas voltadas para área da educação.

Posto esse desafio, Sansone et al (2012) esclarecem que para criar um ecossistema de dados abertos interoperáveis exigir-se-á alavancar diversos esforços. Portanto, o uso das novas tecnologias representa um papel importante ao possibilitarem a integração dos dados, facilitando sua exploração e o reuso da informação.

No levantamento realizado por Santos, Ferreira e Miranda (2017), foi identificado que os principais assuntos abordados com dados abertos educacionais estão voltados a: mineração de dados com foco na área educacional; análise qualitativa dos dados abertos educacionais; teoria sobre o uso de dados educacionais; sistema de apoio a aprendizagem e visualização de dados abertos educacionais.

Berends et al (2017) indicaram que os usuários que consomem dados abertos têm dificuldades para encontrar os dados que estão procurando. Os dados abertos educacionais não fogem dessa realidade. Alcantara et al (2015) apontam que os dados educacionais brasileiros possuem problemas de completude, qualidade da informação e formatos de dados inadequados para reuso. Além disso, os autores mencionam a falta de comunicação entre repositórios. Para se atingir o princípio da rápida integração dos dados abertos, tornando-os conectados, é preciso seguir princípios e padrões que permitam o uso de técnicas e ferramentas que possibilitem a mescla de informações distintas em diferentes contextos.

Ahmed et al (2017) definem Big Data como uma plataforma capaz de consumir diferentes tipos de fontes, assim, facilitando o processo de integração. Williamson (2015) trata como a tecnologia é capaz de coletar grandes quantidades de dados oriundos de várias fontes.

Segundo Miloslavskaya e Tolstoy (2016) e Varpio et al (2020), a governança em um ambiente de dados por meio da construção de um framework permite contextualizar por que determinada solução é relevante.

Desse modo, fica evidenciada a necessidade do desenvolvimento de uma estrutura conceitual que se baseie em estudos existentes, visando explorar o que já é de conhecimento comum e que permita identificar as entradas e saídas em um processo de coleta e apropriação da informação. Portanto, o objetivo deste trabalho é apresentar a proposta de um *framework* conceitual para

aquisição e enriquecimento de dados abertos governamentais de interesse da área educacional de forma que possam ser usados de modo conectado entre si e com outros domínios de dados

2. Referencial Teórico

2.1. Dados Abertos

Dados são considerados abertos quando são livremente acessados, usados, modificados e compartilhados para qualquer propósito, sendo necessário apenas creditar sua autoria e compartilhá-los usando a mesma licença (Rautenberg et al, 2018). Dados Abertos são dados que estão livremente disponíveis para todos utilizarem e redistribuírem como desejarem, sem restrição de licenças, patentes ou mecanismos de controle.

A disponibilização de dados tem se tornado mais comum e ocorre em diferentes contextos. Por outro lado, o relatório de Berends et al (2017) aponta que a exploração de dados abertos ainda encontra barreiras políticas, organizacionais, legais, técnicas e financeiras. A dificuldade se estende também a fatores culturais e das necessidades especificas de cada localidade.

Os dados abertos governamentais foram definidos por Ubaldi (2013), como o movimento que apoia as informações de domínio público que podem beneficiar a sociedade, criando novos produtos e serviços, dando condições para a inclusão social e uma democracia mais participativa. Importante perceber esse movimento como um agregador de iniciativas governamentais ao redor do mundo por mais transparência e *accountability*.

No âmbito do governo federal brasileiro, foram criados diversos dispositivos no intuito de estabelecer e regulamentar a disponibilização de dados abertos governamentais. Destacam-se a Lei de Acesso à Informação nº 12.527/201, Decreto nº 8.777/2016, a Instrução Normativa INDA - IN 4/2012 e o Plano de Dados Abertos (PDA).

O Portal Brasileiro de Dados Abertos (dados.gov.br) é uma das materializações alcançada pelos dispositivos acima citados, ele compõe um dos projetos para oferecer facilidades no acesso as bases de dados públicas. O Portal menciona oito princípios de dados abertos governamentais para ressaltar a importância para a democracia: 1) completo; 2) primários; 3) atuais; 4) acessíveis; 5) compreensíveis por máquinas; 6) não discriminatório; 7) não proprietários; 8) livres de licença (Dados Abertos, 2017).

Em 2011, a W3C iniciou um Grupo de Trabalho de Dados Abertos para orientar as iniciativas governamentais nessa área. Essa iniciativa possibilitou a disponibilização de diversos produtos que servem de guias para os governos que querem abrir seus dados para a sociedade dentre eles destacam-se: para gestores públicos, o Manual dos Dados Abertos: governo (W3C Brasil, 2011a), oferece orientações técnicas e políticas sobre os dados abertos de governo. Com texto voltado para a equipe técnica de desenvolvimento, o Manual dos Dados Abertos: desenvolvedores (W3C Brasil, 2011b), oferece um material mais técnico voltado a orientar na implementação de iniciativas em dados abertos.

Conforme Rautenberg et al (2018), diversos autores propuseram ciclos de vida diferentes para publicação de dados. Attard et al (2015) apresentaram uma adaptação para ciclo de vida dos dados abertos governamentais elucidando as atividades que servem para sustentar esse conceito. A fase de pré-processamento envolve criação de dados (Data Creation), seleção de dados (Data Selection), combinação de dados (Data harmonization) e publicação (Data

Publishing). Após a informação ser disponibilizada entra na fase de exploração com interligação de dados (Data Interlinking), descoberta (Data Discovery), exploração (Data Exploration) e compreensão de como esses dados podem ser aproveitados (Data Exploitation). Finalizando na fase de manutenção (Data Curation).

2.2. *Linked Data*

Zaidan e Bax (2013) afirmaram que a publicação e o consumo de dados na web tendem a ser facilitados com o uso de tecnologias semânticas. Definido por Berners-Lee (2006), o conceito de *Linked Data*, traduzido livremente por dados vinculados, se refere a uma web que pode ser relacionada por meio de dados, facilitando a exploração da informação, para Berners-Lee, essa web deve ser construída aplicando o já conhecido conceito de hipertexto praticado na internet com a diferença de usar dados para referenciar, nessa construção, eles devem ser tratados como documentos. Dessa forma, a partir de um dado uma pessoa ou máquina pode encontrar outros dados relacionados.

Nesse mesmo artigo, Berners-Lee define quatro regras para uma web de dados vinculados. São elas:
1. a primeira regra, usar *Uniform Resource Identifier* - *URI*s[1] para identificar os objetos;
2. a segunda regra, aplicar *links* HTTP sobre as *URI*s;
3. a terceira regra, uso de padrões ou ontologias para organização dos dados;
4. a quarta regra, incluir o maior número possível de *links* para outras *URI*s para a descoberta de mais informações.

Linked Data é um subconjunto da web semântica que usa os conceitos da internet para descrever o mundo real como: identificar fatos, conceitos, pessoas, lugares e fenômenos (Science International, 2015). Dessa forma, esse modelo possibilita a interligação de diferentes conjuntos de dados.

Desde 2007, a W3C é a responsável mundial pelo projeto comunitário aberto voltado à publicação de vários conjuntos de dados interligados chamado de Linked Open Data - LOD (Zaidan & Bax, 2013). Na percepção de Mockus e Palmirani (2015), o conceito de LOD permite o desenvolvimento de um framework propício para o reuso dos dados. São exemplos de aplicações que utilizam desse entendimento: Data.gov.uk; Geonames; US Census; DailyMed entre outros. A Figura 1 mostra o ciclo de vida genérico para dados abertos conectados proposto por Auer (2011).

2.3. Formatos e Licenças

Os formatos de arquivo são acordados e padronizados pela maneira como as informações são codificadas para uso por computadores. Um formato descreve como o arquivo é representado quando armazenado (Geothink, 2020). Com esse entendimento, um formato aberto especifica os arquivos de domínio público. Enquanto que os formatos que são mantidos por empresas privadas são designados como proprietários como DOC, XLSX e XLS.

Dentro desse universo de formato de dados existem a tipificação daqueles que são estruturados e desestruturados. Hurwitz et al (2016) definem dados não estruturados ou apenas

[1] URI - https://www.w3.org/TR/uri-clarification/

desestruturados como aqueles que não possuem uma estrutura definida, ao contrário dos estruturados, podendo conter diversas formas em seu conteúdo.

> Para ser considerado um dado aberto, o conjunto de dados deve estar disponível em um formato de especificação aberta, não proprietário, e estruturado, ou seja, que possibilite seu uso irrestrito e automatizado através da Web. Além disso, é imprescindível que seja utilizado um formato amplamente conhecido (Brasil, 2012).

Figura 1 - Ciclo de vida do *Linked Open Data* (Auer, 2014).

De acordo com o Decreto n 8.777 (Brasil, 2016a), os dados abertos disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte. No Quadro 1, apresenta as principais licenças abertas que podem ser utilizadas na disponibilização de dados abertos (Rautenberg et al, 2018).

2.4. Esquema Cinco Estrelas

O método proposto por Berners-Lee (2006) classifica a qualidade de um conjunto de dado entre 1 a 5 estrelas conforme um conjunto de características que iniciam em tornar as bases disponíveis na Web, elevando para duas estrelas quando os são estruturados, três quando soma-se que são não proprietários. A quarta estrela significa utilizar URIs para identificar recursos, por fim, a quinta estrela significa conectar seus dados com outras fontes de dados.

Quadro 1 – Licenças abertas

Licenças	**Descrição**
Licença Comum Criativa (Creative Commons Licenses - CC)	Permite a cópia, reuso, distribuição e modificação do dado original. Destacam-se alguns subconjuntos de licenças: atribuição (BY)
Dados Abertos Comuns (Open Data Commons Attribuittion Licence - ODC)	licença especifica para dados e banco de dados abertos. Com três derivações: ODC-ODbL; ODC-BY e ODC-PDDL
Licença de Governo Aberta (The Open Government Licence - OGL)	licença voltada para que os entes públicos possam copiar, distribuir, adaptar os dados, podendo o seu uso de forma comercial ou não.

Fonte Rautenberg *et al* (2018)

A classificação proposta na Figura 2 pode ser entendida como uma escada evolutiva para interligação de dados. Para chegar ao degrau superior depende antes de passar pelo degrau inferior. Rautenberg et al (2018:19) descreveram que a partir de uma estrela, quando os conjuntos de dados têm seus recursos disponíveis por uma licença aberta (Open License - OL), não importando o formato, garantindo que usuários são livres para utilizar os dados com qualquer finalidade, podendo até modificá-los, adaptá-los e distribuí-los; duas estrelas, quando tem seus recursos disponíveis como dados estruturados sendo legíveis por máquinas (Readable Machine - RE), por exemplo, planilha no lugar de imagem escaneada; três estrelas, quando utiliza formatos não proprietários (Open Format - OF), por exemplo, arquivo CSV e não planilha do tipo XLSX; quatro estrelas, quando utiliza Identificadores de Recursos Uniformes (Uniform Resource Identifier - URI) para identificar os dados, possibilitando a criação de links entre eles de forma a facilitar o reuso. Até os metadados são publicados em forma de URIs; e cinco estrelas, quando conjuntos de dados contém links para outros conjuntos de dados relacionados ao mesmo tema, configurando os dados conectados ou Linked Data – LD.

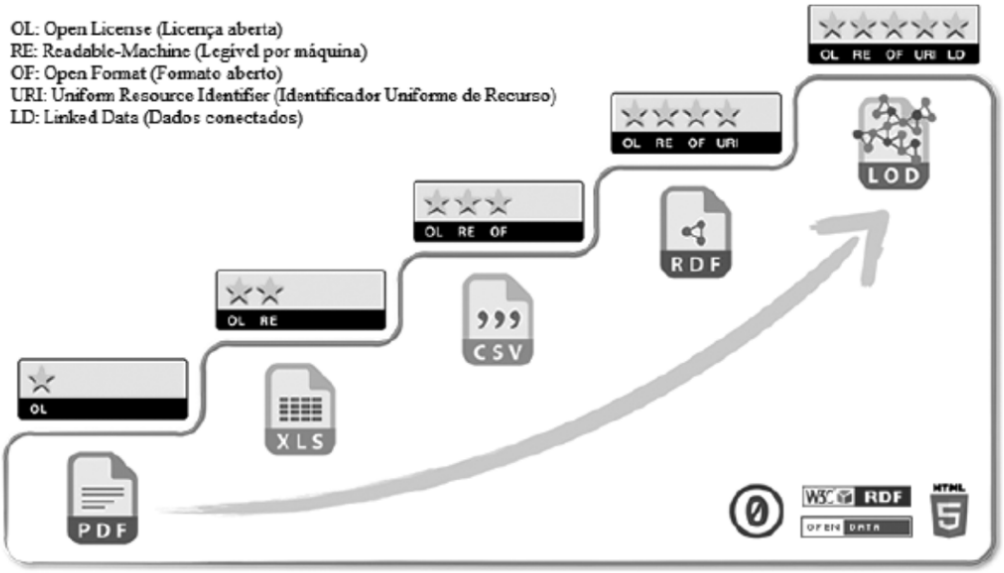


Figura 2 - Esquema cinco estrelas (Rautenberg et al, 2018:19).

2.5. Modelo de interoperabilidade do Governo Brasileiro

De acordo com o Decreto nº 8.777 (Brasil, 2016a), os dados abertos devem ser disponibilizados na internet tornando-se acessíveis e estruturados em formato aberto possibilitando o processamento por máquina. A arquitetura de Padrões de Interoperabilidade de Governo Eletrônico, a ePING define um conjunto mínimo de premissas, políticas e especificações técnicas que regulamentam a utilização da Tecnologia de Informação e Comunicação (TIC) na interoperabilidade de serviços de Governo Eletrônico (Brasil, 2014, art. 1º).

Portanto, a ePING estabelece as condições de interação com os demais Poderes e esferas de governo e com a sociedade em geral. Ela traz também princípios e orientações para a configuração das TICs dentro do governo. Segundo o Documento de Referência da ePING (Brasil, 2018a), a arquitetura propõe questões organizacionais e semânticas, tratando todo o ciclo da informação no governo federal.

A Infraestrutura Nacional de Dados Abertos - INDA é um 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, em conformidade com o disposto na ePING (Brasil, 2020d).

A Infraestrutura Nacional de Dados Espaciais – INDE foi instituída pelo Decreto nº 6.666 de 2008. A INDE é um conjunto integrado de tecnologias; políticas; mecanismos e procedimentos de coordenação e monitoramento; padrões e acordos, necessário para facilitar e ordenar a geração, o armazenamento, o acesso, o compartilhamento, a disseminação e o uso dos dados geoespaciais de origem federal, estadual, distrital e municipal (Brasil, 2020e).

A lei nº 12.527, conhecida como Lei de Acesso a Informação - LAI trouxe a necessidade de evoluir a antiga Lista de Assuntos de Governo – LAG, que estabelecia rol de termos usados no governo federal para uma lista que contemplava todos os assuntos relacionados com a atuação do governo. Em 2010, surge o VCGE – Vocabulário Controlado e Governo Eletrônico (Brasil, 2016b). O VCGE é um vocabulário controlado para indexar informações (documentos, bases de dados, sites, etc.) para o governo federal. Ele foi projetado com dois objetivos básicos: ser usado como interface de comunicação com o cidadão; e ferramenta de gestão.

Apesar do escopo legislativo ser no âmbito federal, os normativos apresentam recomendações gerais apoiadas em boas práticas de disponibilização de dados abertos governamentais almejando a interoperabilidade desses conjuntos. Logo, é esperado que os dados disponibilizados por estados e municípios sigam os mesmos princípios para que as bases de dados possam trabalhar em conjunto.

2.6. *Data Lake*

Data Lake é uma coleção de instâncias de armazenamento de vários ativos de dados adicionais às origens de dados de origem. Esses dados são armazenados como uma cópia quase exata ou quase exata do formato de origem (Gartner, 2020). Evoluindo o entendimento sobre o tema, data lake pode ser entendido como uma estratégia de ser um repositório para armazenar os dados deixando-os disponíveis para serem trabalhados de acordo com as necessidades de análise.

A implantação de um data lake geralmente usa segregação lógica separando o ambiente em uma estrutura em camadas. Pasupuleti e Purra (2015, p.10) apoiam esse entendimento que se trata de uma solução complexa que envolve várias camadas. Da mesma forma, Miloslavskaya e Tolstoy (2016) comentaram que pode ser dividido em três camadas: uma para os dados brutos (raw data), outra para receber os dados e uma terceira para armazenar informação.

O desenho da arquitetura de um *data lake* pode abranger diversas tecnologias e ambientes, permitindo a interoperabilidade entre diversos componentes. Ao mencionarem as características desse tipo de ambiente, Miloslavskaya e Tolstoy (2016) destacaram: 1) uma estrutura escalável para concentrar os dados permitindo o processamento e armazenamento dos dados conforme a necessidade dos usuários; e 2) a centralização de um catálogo para indexação das fontes de informação o que agilizaria a localização das bases de dados.

Khine e Wang (2018) se referem ao *data lake* como uma fase inicial de um ambiente propício para *big data*. Para Pasupuleti e Purra (2015:12), é um ambiente que torna possível armazenar os dados encontrando padrões entre eles. A criação de um *data lake* onde todas as informações são capturadas e armazenadas em seu formato original. Posteriormente, conforme a necessidade, esses dados podem ser transformados para serem processados e analisados.

Um *data lake* pode suportar o tratamento sobre os dados permitindo análises e usos operacionais dos dados. Portanto, após o estabelecimento de um fluxo de consumo de dados (pipeline) no *data lake*, é preciso agregar características como velocidade para o consumo de dados. Outra característica interessante, é a capacidade de processamento de um volume elevado de arquivos de dados que serão ingeridas. Acrescenta-se a variedade de formatos que esses dados estão espalhados pela internet.

2.7. *Big Data*

A grande quantidade de informações geradas pela sociedade contemporânea, proveniente de inúmeras fontes de dados como a web, redes sociais e dispositivos móveis, e as formas automatizadas a partir das quais essas são coletadas, promove um grande volume de dados nos sistemas computadorizados (Machado, 2017). *Big Data* não é uma única tecnologia, mas uma combinação de tecnologias novas e antigas que ajudam as empresas conseguirem ideias viáveis" (Hurwitz et al, 2016:14).

Soluções em *big data* surgiram como uma solução para análise de grandes volumes de dados. Machado (2017:15-17) divide as características dessa abordagem em dois grupos de "V's": os nativos que são volume, a variedade e a velocidade; e os os adquiridos, que são a variabilidade, a veracidade e o valor:
- Volume: apesar de indicar relação com tamanho dos conjuntos de dados. Essa característica também está relacionada ao tempo de comunicação entre as bases de dados e o ambiente de análise;
- Variedade: trata-se de um ambiente que recebe os mais diversos tipos e estruturas de dados que podem sofrer constantes atualizações;
- Velocidade: quanto mais rápida for a relação entre obtenção e análise sobre o volume de dados, melhor;
- Variabilidade: relativa às medidas de dispersão de um conjunto de dados;
- Veracidade: relacionada a confiabilidade sobre a origem dos dados;
- Valor: se refere a abordagem que será feita sobre os dados para que eles agreguem valor.

Conforme explicado por Machado (2017:25), dentro de um ambiente big data "...uma determinada aplicação pode ter ênfases diferentes em cada um dos "V's"". Ao implementar essa abordagem de controle de dados, deve-se primeiro pensar em qual o propósito pretendido e depois analisar quais deles é o mais importante e dessa forma balancear essas características dentro da solução big data.

Um dos aspectos mais relevantes de ambientes big data é a capacidade de tratar com uma variedade de dados. A habilidade de trabalhar com conjuntos de dados estruturados, aqueles que já possuem uma estrutura padronizada para leitura ou extração em formato de linhas e colunas. Como também a capacidade de trabalhar com informações não estruturadas como arquivos de vídeo, documentos e imagens. E o melhor, fazendo ambas as estruturas se combinarem de alguma forma.

Dumbill (2014) afirma que, inicialmente, as aplicações big data foram desenvolvidas com foco apenas nas necessidades da área de tecnologia da informação (TI) agregando pouco valor para a tomada de decisão do negócio. Esse movimento inicial, acarretou na criação de silos de dados que pouco ajudavam no atendimento do negócio, uma vez que a obtenção de valor comercial depende de um dispendioso esforço de extração.

2.8. *Big Open Linked Data* (BOLD)

O *linked data* trouxe um novo caminho para a integração de dados na *web*. Permitindo a criação de valor entre diferentes tipos de conjuntos de dados. De um lado, temos pesquisadores e demais interessados e com a necessidade de lidar com os conhecidos "V's" do conceito de big data: volumes de dados; necessidade de alta velocidade de resposta e a variedade em dados.

Hitzler e Janowicz (2013) e Janssen e Kuk (2016), ao comentarem sobre o quarto paradigma da Ciência, enfatizaram que a estratégia de exploração de dados é mais um novo campo da ciência e pesquisa de dados, no qual a transdisciplinaridade entre pessoas e domínios pode acontecer graças à abordagem de *linked data* agregada à tecnologia de *big data*, possibilitando novos *insights* por meio do uso de *datasets* que já existem. Com esse pensamento, Janssen e Hoven (2015) conceituam a sigla BOLD como integração de três grandes desenvolvimentos que afetam a sociedade. Trata-se de um acrônimo usado para representar o uso de dados na era digital, referindo-se a natureza mutável dos dados. Segundo Janssen et al (2017), o Big Open Linked Data desempenha um papel central na coleta, transformação e na combinação e compartilhamento de dados originados de várias fontes.

2.9. *Framework* Conceitual

Na área da tecnologia da informação, frameworks são relacionados a estruturas pré-concebidas, que conseguem fornecer funcionalidades em comum dentro de um domínio. Johson (1997) caracteriza framework como um conjunto de classes que incorpora um projeto abstrato que provê uma ou mais soluções para um conjunto de problemas relacionados.

Comumente, *framework* é usado para explicação de um problema baseado em conceitos relacionados. Ele ajuda a entender desde as intenções e expectativas de uma pesquisa. Uma descrição de uma estrutura que contribua no reporte de uma pesquisa de duas maneiras: (1) identifica variáveis de pesquisa e (2) esclarece as relações entre as variáveis. Vinculado ao problema da pesquisa, o *framework* conceitual prepara o cenário para a apresentação da questão de pesquisa e impulsiona a investigação que está sendo realizada (McGahie, Bordage, & Shea, 2001).

O *framework* conceitual incorpora o paradigma da pesquisa especificando a direção pela qual a pesquisa terá que ser realizada. Trata-se da ideia do pesquisador sobre como o problema de pesquisa terá que ser explorado. Regoniel (2016) entende *framework* conceitual como a representação de um fenômeno concebido por meio de um apanhado sobre a literatura existente e sobre determinado fenômeno. Complementando, Imenda (2014) o define como um resultado final a partir da reunião de diversos conceitos relacionados. Pontua ainda esse arcabouço conceitual que permite: a) explicar ou prever um determinado evento; b) prover uma compreensão de um problema de pesquisa. Trata-se do processo de conceber uma estrutura conceitual é semelhante a um processo indutivo no qual pequenas peças individuais (neste caso, conceitos) são unidas para formar um mapa de possíveis relacionamentos.

Swaen (2020) propôs alguns passos para a formulação de uma estrutura conceitual: (1) identificar a ideia ou paradigma do framework conceitual; (2) identificar as variáveis; (3) apontar as variáveis dependentes e independentes; e (4) projetar a estrutura conceitual. Para realizar essas atividades deve-se adotar procedimentos metodológicos para criação da estrutura conceitual.

Com o exposto, entende-se que o conhecimento repassado por Regoniel (2016) permite a criação da estrutura conceitual devendo ser compreendido em conjunto com os processos expostos por Swaen (2020) e McGahie, Bordage e Shea (2001), que aprofundam na identificação e no relacionamento das variáveis.

2.10. Síntese

Com essa breve revisão, é possível perceber que para a criação de uma estrutura voltada para big data, identifica-se 3 elementos principais: 1) captura de dados e pré-processamento; 2) processamento e armazenamento de dados e; 3) interação com os dados. No caso especifico dos dados abertos, Casaes (2019) propôs *framework* que agrega elementos voltados a um processo de gestão e governança direcionado a dados abertos.

Desse modo, o estabelecimento de estruturas que sejam capazes de auxiliar os produtores de dados nos processos de coleta e preparo da informação são uma das preocupações a serem sanadas antes da proposição do uso de tecnologias e padrões. Com intuito de prover ao usuário final a possibilidade de se beneficiar com o grande volume de dados produzidos.

3. Metodologia da Pesquisa

Inicialmente a pesquisa foi realizada de forma dedutiva para analisar por meio da revisão de literatura e do levantamento teórico sobre o tema dessa pesquisa. Posteriormente, o trabalho assumiu um caráter indutivo por considerar a pesquisa empírica, com as observações percebidas pelo pesquisador identificados na fase de coleta que permitem a proposição de um *framework* conceitual. As pesquisas descritivas têm como objetivo primordial a exposição das características de determinada população e fenômeno ou o estabelecimento de relações entre variáveis (Gil, 2002). Dessa forma, determinar a natureza da relação entre as variáveis identificadas.

No que tange a investigação, a pesquisa bibliográfica objetivou a criação de um arcabouço teórico a partir do levantamento da literatura relacionada para apurar o conhecimento acerca do tema desse estudo. A pesquisa documental utilizou como fonte de informação os Planos de Dados Abertos (PDA) de cada ente público mencionado com a finalidade de identificar e entender as bases de dados abertos disponibilizadas.

As informações levantadas permitiram a elaboração preliminar de uma estrutura conceitual para o aproveitamento de dados abertos educacionais entre os entes públicos. Essa construção preliminar serviu na identificação de possíveis variáveis que influenciam no desenvolvimento dessa estrutura. A continuidade do estudo foi o de coletar e validar as informações até aqui desenvolvidas com grupos de interesse em dados abertos educacionais. Por fim, após a coleta e análise de todas as informações obtidas permitiram a construção final do framework conceitual.

3.1. Revisão documental

As principais informações disponibilizadas em dados abertos sobre a educação brasileira são produzidas pelas principais entidades federais desse setor como Ministério da Educação – MEC, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – Inep, Fundo Nacional de Desenvolvimento da Educação – Fnde e Coordenação de Aperfeiçoamento de

Pessoal de Nível Superior – Capes. Essas entidades públicas disponibilizam seus Planos de Dados Abertos (PDA).

Conforme o Quadro 3, a investigação partiu da análise dos PDA publicados, no momento da realização dessa pesquisa, para caracterização e entendimento das bases de dados abertos disponibilizadas pelos entes educacionais.

Quadro 3 - Planos de Dados Abertos do estudo

PDA	Documento
MEC (BRASIL, 2020b)	Plano de Dados Abertos do MEC para o biênio 2020/2021.
Inep (BRASIL, 2020a)	Plano de Dados Abertos do Inep para o biênio de 2020/2021.
Fnde (BRASIL, 2018)	Plano de Dados Abertos do Fnde para o biênio 2018/2019.
Capes (BRASIL, 2017)	Plano de Dados Abertos – Capes - triênio 2017 a 2019.

A pesquisa sobre os documentos procurou identificar se os documentos atendem ao Manual de Elaboração de um PDA (BRASIL, 2020c), quais as bases de dados que são disponibilizadas de forma aberta, a descrição de cada uma e a quantidade de produtos disponibilizados.

A caracterização das bases de dados foi materializada em um inventário. O levantamento agrupou as informações relacionadas ao conteúdo como: área de interesse, unidade de análise, abrangência geográfica e a temporalidade de atualização das mesmas. Outra parte da análise consistiu em identificar os formatos que as bases de dados e seu metadados são disponibilizados verificando se alguma licença de uso dos dados estava associada.

A partir das informações coletadas realizou-se uma análise qualitativa por meio da classificação de Cinco Estrelas proposta por Berners-Lee (2006). Essa categorização consiste em atestar o nível de conectividade das bases de dados abertos disponibilizadas.

3.2. Entrevistas

A presente pesquisa utilizou-se da entrevista na modalidade semiestruturada. Segundo Gil (2002), é aquela que possui maior maleabilidade, cabendo ao entrevistador guiar o entrevistado durante a aplicação das questões. Anteriormente à elaboração das questões, os PDA dos entes educacionais haviam sido estudados, bem como as principais características das bases de dados abertos disponibilizada por esses órgãos. Dessa forma, embasaram a construção do framework conceitual preliminar, juntamente com a revisão de literatura.

A coleta dos dados ocorreu em dois momentos. O primeiro, foi a realização de entrevistas com os especialistas em dados abertos do INEP. O segundo momento ocorreu a transcrição e verificação das informações coletadas. O roteiro de entrevista aplicado juntamente com o framework conceitual preliminar contém questões que foram construídas no intuito de deixar o entrevistado a possibilidade de discorrer sobre o tema proposto. Os procedimentos adotados para realização das entrevistas seguiram os seguintes critérios: construção do instrumento de pesquisa; estabelecimento de um protocolo de entrevista; roteiro de pesquisa aplicado nas entrevistas com os especialistas; envio do convite para os especialistas com a contextualização do assunto da pesquisa; agendamento e realização das entrevistas; e transcrição e análise dos dados.

As entrevistas com os especialistas apresentadas no Quadro 4 ocorreram em outubro de 2020 por videoconferências, com profissionais com experiência em coleta, tratamento e disponibilização das informações, principalmente, nos produtos relacionados a dados abertos do Inep. Todas as entrevistas foram gravadas utilizando a ferramenta Google Meets.

Quadro 4 - Síntese das entrevistas com os especialistas

Data	Identificação	Área	Tempo
16/10/2020	Entrevistado A	Coordenação de Sistematização e Disseminação da Informação	00:48’15’’
20/10/2020	Entrevistado B	Coordenação de Controle de Qualidade e Tratamento da Informação	01:24’15’’
22/10/2020	Entrevistado C	Coordenação de Instrumentos e Medidas	00:38’11’’
23/10/2020	Entrevistado D	Coordenação de Sistemas de Informação	00:13’14’’
27/10/2020	Entrevistado E	Projetos de Business Inteligence	00:22’50’’

Fonte: elaborado pelos autores.

4. Resultados da Pesquisa

4.1. Investigação sobre os dados abertos educacionais

O Plano de Dados Abertos (PDA) é o instrumento que operacionaliza a Política de Dados Abertos do Poder Executivo Federal, pois planeja as ações que visam a abertura e sustentação de dados nas organizações públicas. Cada órgão/entidade possui a obrigação de elaborar um PDA com vigência de dois anos, a contar da publicação do documento (BRASIL, 2020c, p. 05)

No Brasil, para garantir a publicidade, transparência na administração pública e acesso às informações pela sociedade a publicação do Plano deve estar alinhada, à Política de Dados Abertos, instituída pelo Decreto nº 8.777/2016, à Lei nº 12.527/2011 e aos Planos de Ação Nacional sobre Governo Aberto.
Com relação a legalidade, os documentos examinados são amparados pela Lei Complementar nº 101/2000, Decreto nº 6.666/2008, Decreto nº 15/2011, Instrução Normativa nº4/2012, Plano de Ação da INDA, Plano de Ação Nacional Governo Aberto, Lei nº 12.527/2011, Portaria nº 03/2007, Decreto nº 8.638/2016, Decreto nº 8.777/2016.

Dentre as diretrizes gerais definidas nos documentos analisados estão a adoção dos padrões estabelecidos pela ePING, pelo Governo Eletrônico, pela Infraestrutura Nacional de Dados Abertos – INDA e pela Infraestrutura Nacional de Dados Espaciais – INDE.

No que se refere à denominada motivação social dos documentos analisados demonstrou similaridades em questões para promover a transparência, ampliar a participação social, além de ofertar subsídios para o desenvolvimento de novas tecnologias para auxiliar à tomada de decisão por meio do compartilhamento de dados e assim servir como instrumento de melhoria na gestão pública. O Quadro 5 apresenta a totalidade de produtos trazidas pelas bases de dados abertos governamentais que são disponibilizadas.

Com base na análise dos documentos, foi possível entender a metodologia adotada para a disponibilização das bases de dados ocorrem: como são identificadas e priorizadas. Por seguirem modelos legais em comum, o fluxo de processos para a disponibilização dos planos de dados abertos e do seu conteúdo passam pela alta gestão de cada órgão, que por sua vez, coordenam e são municiados pelas áreas tático-operacionais.

Quadro 5 - Plano de Dados Abertos dos entes educacionais

PDA	Documento	Qtd. de arquivos de dados
MEC (BRASIL, 2020b)	Plano de Dados Abertos do MEC para o biênio 2020/2021.	22 arquivos
Inep (BRASIL, 2020a)	Plano de Dados Abertos do Inep para o biênio de 2020/2021.	87 arquivos
Fnde (BRASIL, 2018b)	Plano de Dados Abertos do Fnde para o biênio 2018/2019.	27 arquivos
Capes (BRASIL, 2017)	Plano de Dados Abertos – Capes - triênio 2017 a 2019.	28 arquivos

Fonte: elaborado pelos autores.

Com a análise sobre o descritivo cada base de dados foi possível vislumbrar informações relacionadas e de interesse, porém por se tratar de entidades públicas da área educacional que são vinculadas ao um mesmo ente vinculante foi percebido a falta de uma seção em comum entre os planos na qual relacionariam esses dados.

4.2. Caracterização das bases de dados abertos

Para a caracterização das bases de dados tomou-se os planos de dados abertos e os *sites* institucionais de cada ente público da área educacional, mencionados nessa pesquisa. Esse procedimento permitiu identificar e caracterizar as bases de dados abertos e os produtos de cada uma. Lembrando que cada base de dados mencionadas nos documentos analisados são compostas por um ou mais arquivos de dados associados. Na prática, uma base de dados pode ser formada por um conjunto de arquivos de dados.

O levantamento mostrou que a maior parte dos conjuntos de dados disponíveis é de interesse da educação básica com 101 conjuntos de dados, seguido da educação superior com 26 conjuntos e 21 sobre a pós-graduação. O ensino técnico sozinho soma 7 conjuntos de dados. A evidente diferença entre o quantitativo de dados disponibilizados para educação básica pode ser justificável devido ao tamanho e à complexidade que essa área possui. Por outro lado, essa desproporção com as demais áreas pode significar a ausência de informação.

A pesquisa permitiu elencar as principais unidades de analise abrangidas pelos conjuntos de dados. A fim de proporcionar uma melhor consolidação das informações, foi adotado o procedimento de padronizar a categorização dos termos para designar as unidades de análise identificadas. O ranking apresenta uma tríade de termos dominante composta pelas Instituições de Ensino[2] com 21%, Estudantes, 18%, e Professores 9%.

Com relação à abrangência geográfica dos conjuntos de dados abertos, o destaque fica para informações a nível Municipal com 60%, Instituições de Ensino com 12% e nível Estadual com 7%. Evidencia-se que a maioria dos dados disponibilizados chegam até o nível de município, alguns deles até nível de espaço físico quando se refere a instituições de ensino.

A tempestividade de atualização dos conjuntos de dados apresentou com destaque a granularidade Anual com 57%; Mensal 13%; Semestral com 7%. Os números apontam que mais da metade é atualizada anualmente. Observando as atualizações mensais e semestrais

[2] Instituições de Ensino nessa pesquisa representam tanto as Escolas de Educação Básica como os Instituições de Ensino Superior.

juntas representam cerca de 20% dos dados, portanto esses são renovados mais de uma vez dentro de um mesmo ciclo.

Passada essa primeira parte da análise, a investigação continuou na identificação e categorização dos formatos dos arquivos de dados disponibilizados, a mesma linha de análise foi adotada na identificação dos metadados dos dados abertos. Acrescenta-se que foram verificados se estes conjuntos de dados possuem alguma licença atribuída.

A predominância do formato CSV (média 78%), sendo praticamente um padrão adotado entre os entes governamentais publicadores. A disponibilidade de dados em formatos interoperáveis por máquina como CSV muito se deve pela obediência as diretrizes estabelecidas pela ePING. Entretanto, nota-se baixa adesão a formatos mais recentes como RDF e URI ambas com 11%. O formato XML usado apenas por Fnde (70%) e MEC (9%) e JSON usado apenas pelo Fnde (48%).

Com relação aos metadados analisados, quando disponíveis, a concentração de arquivos no formato PDF e no formato de planilha (XLS, XLSX, ODS). Analises sobre os formatos (dados e metadados) apresenta uma tendência de que quando um padrão de tipo de arquivo é adotado, dificilmente o publicador adote outro formato, mantendo sempre o mesmo. Percebe-se que o ente publicador pode divulgar algumas de suas bases de dados em formato voltado a dados conectados como RDF, mas acaba por adotar um padrão que dificulta a conectividade, no caso metadados.

A análise dos planos de dados e dos conjuntos de dados abertos procurou identificar: 1) se era especificada alguma licença de dados abertos; 2) em caso de uso de uma licença, qual a licença utilizada. Os planos de dados abertos analisados não foram claros em especificar se as bases de dados obedecem a alguma licença. Apenas se limitam a mencionar que seguem os princípios de publicação dos dados abertos e que os dados serão divulgados sob licença aberta. A Instrução Normativa Nº 4 (INDA) é clara ao assinalar em seu art. 6º que compete ao Comitê Gestor de cada órgão, definir o modelo de licença para os dados abertos. Algumas informações sobre licenças atribuídas foram verificadas por meio das informações que por ventura alguns dos portais institucionais traziam.

A maioria dos dados abertos publicados, cerca de 66% dos arquivos, não possuem uma licença especifica. O restante está sob a licença CC-BY com 28% e 5% registradas com ODbL. A escolha de um tipo de licenciamento permite uma maior liberdade para o uso dos dados. Ainda, de acordo com o Decreto n 8.777 (Brasil, 2016a), os dados abertos disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte.

Realizou-se uma análise qualitativa por meio da classificação de Cinco Estrelas proposta por Berners-Lee (2006). O método aplicado classifica a qualidade de um conjunto de dado entre 1 a 5 estrelas o que significa o quanto esses dados são acessíveis para pessoas e máquinas observando também o grau de conectividade entre eles.

O resultado da aplicação desses critérios ficou resumido em 13 arquivos sem estrelas; 01 com duas estrelas; 96 com duas estrelas; seguindo de 50 arquivos com três estrelas; 03 com quatro estrelas; finalizando com nenhum arquivo classificado com cinco estrelas. A classificação foi realizada sobre os arquivos de dados disponíveis. Os critérios foram ponderados com alguns entendimentos: os planos de dados abertos apenas identificam que os dados publicados atendem

aos oito princípios de dados abertos sem especificar a qual licença esse dado realmente pertence. Reforçando o entendimento sobre as licenças, Rautenberg et al (2018, p. 19), entendem que os conjuntos de dados têm seus recursos disponíveis por uma licença aberta garantindo que usuários são livres para baixa-los e utiliza-los com qualquer finalidade, podendo até modificá-los, adaptá-los e distribuí-los.

A designação feita pelo Esquema Cinco Estrelas de que o dado seja publicado sob uma licença aberta (*Open License*). Dessa forma, a analise realizada considerou 3 ou 4 estrelas aqueles arquivos que possuem de fato um tipo de licença aberta especificada e que atendam ao restante dos critérios estabelecidos pelo esquema. A classificação continuou sobre os demais arquivos que apesar de terem sido publicados em formatos legíveis por máquina como CSV, JSON ou XML, mas sem a designação de uma licença aberta optou-se pela classificação de 2 estrelas, da mesma forma, as bases disponibilizadas apenas no formato de planilha (XLS, ODS ou XLSX). Alguns conjuntos de dados apesar de citados nos PDA, não foram encontrados ou não estavam acessíveis no momento dessa análise, portanto, não foi possível atribuir nenhuma estrela.

O fichamento dos conjuntos de dados dos principais entes federais responsáveis pela educação brasileira permitiu observar a necessidade de uma evolução com relação ao conceito de dados abertos conectados, uma pequena parte, apenas 3 coleções de dados apresentam a perspectiva do chamado *linked data*, o restante, ressalta-se a maioria, dos dados ainda estão entre 2 ou 3 estrelas sendo essa última o mínimo para considerar um dado aberto que permita conectividade. Com relação a distribuição de dados e a atribuição de cada órgão público, poderia ser adotado a constância de inserir atributos comuns de forma facilitar a conexão entre as bases de dados. Nessa linha, uma uniformidade informacional desde a nomenclatura de variáveis, algo como o uso padronizado de metadados e até de vocabulários em comum facilitaria a interoperabilidade entre eles.

Outra questão percebida que algumas bases de dados são disponibilizadas em formato comprimido, no caso o ZIP. Então qualquer atividade que envolva automatizar a coleta e leitura das fontes de dados deve observar essa necessidade de descompactar esses tipos de arquivos após recepciona-los.

4.3. Combinação de dados que tragam informações úteis

Os indicadores educacionais atribuem valor estatístico à qualidade do ensino, atendo-se não somente ao desempenho dos alunos, mas também ao contexto econômico e social em que as escolas estão inseridas. Eles são úteis principalmente para o monitoramento dos sistemas educacionais, considerando o acesso, a permanência e a aprendizagem de todos os alunos. Dessa forma, contribuem para a criação de políticas públicas voltadas para a melhoria da qualidade da educação e dos serviços oferecidos à sociedade pela escola (Brasil, 2020g).

O Censo Escolar é o principal instrumento de coleta de informações da educação básica e a mais importante pesquisa estatística educacional brasileira (Brasil, 2020f). Os dados do Censo Escolar são a principal referência para a gestão de programas federais, tais como: Programa Nacional do Livro Didático (PNLD), Programa Nacional de Alimentação Escolar (PNAE), Programa Nacional de Apoio ao Transporte Escolar (PNATE), Programa Dinheiro Direto na Escola (PDDE), dentre outros.

Além disso, os resultados obtidos no Censo Escolar servem de base para o estabelecimento de diversos indicadores educacionais, sobre o rendimento (aprovação e reprovação) e movimento

(abandono) escolar dos alunos do ensino Fundamental e Médio, juntamente com outras avaliações do Inep (Saeb e Prova Brasil), são utilizados para o cálculo do Índice de Desenvolvimento da Educação Básica (IDEB).

Adeodato et al (2014) e Ferreira (2016) partiram dos dados do censo para observar o desempenho escolar. O primeiro buscou criar um indicador de qualidade observando aspectos econômico-financeiros. Já Ferreira (2016), ao investigar aspectos da educação infantil (infraestrutura, financiamento público) em uma região metropolitana. Nos dois estudos, os autores perceberam a necessidade de agregar outras variáveis para subsidiar os resultados encontrados. Os dados do censo podem ser enriquecidos a partir de outras fontes de informação.

Muniz e Carvalho (2007) destacaram que um dos objetivos da PNAE é o de contribuir com a redução dos índices de evasão escolar com a formação de hábitos alimentares e no aumento da aprendizagem. Informações úteis para compor as informações sobre o aproveitamento e índices de evasão calculados a partir do censo educacional. A base da PNAE disponível é composta por um conjunto de 5 arquivos de dados estruturados. Pela classificação de dados conectados, a base alcançou nível de 3 estrelas por disponibilizar suas informações em formato não proprietário como o CSV compreensivo também para leitura de máquina. Os metadados estão em formato não estruturado PDF.

Na mesma linha, o PDDE traz em seus dados a informação da escola assistida pelo programa. Chudrik (2013) acrescentou que um programa dessa natureza objetiva implementar a infraestrutura física e pedagógica, reforçando a gestão escolar nos planos financeiro, administrativo e didático.

A base alcançou 2 e 3 estrelas em nível de conectividade sendo composta por 6 arquivos de dados disponibilizados nos formatos estruturados (CSV) e semiestruturados (JSON e XML). Todos possuem metadados no formato não estruturado (PDF). Conforme a licença designada, as informações sobre a execução financeira do programa e relação de escolas passíveis de atendimento que permitem a cópia, reuso, distribuição e modificação e até uso comercial.

Os dados da PDDE podem ser relacionados diretamente com os dados sobre as escolas trazido pelo censo da educação. Além disso, os dados desse programa podem compor os indicadores educacionais. Silva (2015) observa que os problemas financeiros nas escolas refletem no desenvolvimento de projetos e atividades e nas avaliações institucionais já que muitas delas demandam não só de ações pedagógicas, mas também como financeiras.

Com esse pensamento, é possível utilizar outras bases de dados a fim de enriquecer ainda mais uma fonte de informação. Com nível 3 estrelas e com licença que permite utilização até com fins comerciais (CC-BY), a possibilidade de conexão dos dados do PNATE tomando por partida as informações trazidas pelo censo educacional para conceder os recursos financeiros necessários para subsidiar o transporte escolar. Os arquivos de dados estão em formatos acessíveis (CSV, JSON e XML). O resultado desse programa educacional pode ser usado nos indicadores educacionais do próprio censo para, por exemplo, melhor entendimento da evasão escolar. Esse tipo de informação poderia demonstrar se as localidades com mais apoio são aquelas que retém os melhores resultados.

Os dados censitários do ensino superior são utilizados de forma articulada com outras políticas públicas como o Exame Nacional de Desempenho de Estudantes (ENADE), o Exame Nacional

do Ensino Médio (ENEM), o Fundo de Financiamento Estudantil (FIES), o Programa Universidade para Todos (PROUNI), o Sistema de Seleção Unificada (SISU) (Brasil, 2015).

Por meio dos dados do censo superior é possível obter algumas informações relacionadas a bolsas. Contudo, a abrangência da rede de ensino superior e a forma como as informações são coletadas acabam sendo variáveis que podem influenciar na exatidão das informações. Por conseguinte, o uso de outras fontes externas, por exemplo, vindas do FNDE (FIES) e MEC (PROUNI) podem auxiliar o Inep no processo de validação das informações inseridas no Censo. Os dois arquivos de dados foram classificados nessa pesquisa com 2 e 3 estrelas respectivamente.

Com o exposto até aqui, percebe-se que o universo informacional para enriquecimento dos dados abertos educacionais é bem amplo. Ressalta-se que apenas os dados abertos disponibilizados entre MEC, INEP, FNDE E CAPES estão sendo considerados. Fato que com o auxílio da tecnologia é possível usar as diversas fontes de dados por meio da sistematização na coleta de dados e na construção informacional para uma melhor gestão do conhecimento.

4.4. Metodologia para extrair informações

Um aspecto importante é como extrair as informações de fontes diferentes. As informações podem ser obtidas por meio da criação de um pipeline ou fluxo de dados. Um *pipeline* é uma conexão entre múltiplos nós que existem para dar suporte ao movimento de dados através de servidores (Hurwitz et al, 2016:110). O *pipeline* inicia por meio de agentes de extração. Para o estudo aqui proposto, essas bases são aquelas advindas dos Planos de Dados Abertos até aqui estudados.

Com esse cenário é importante definir como será o método para extração, transformação e carga das bases de dados. Almeida e Santos (2014) entendem que métodos do tipo ETL (Extract, Transform e Load) trazem simplificação ao processo de migração dos dados, fornecendo formas padronizadas ao processo. Os dados são coletados, são realizadas as transformações necessárias e depois os dados são armazenados. É um padrão adotado por *Business Intelligence* (BI) e *Data Warehouse* (DW). Hurwitz et al (2016:284) definem que métodos ELT (Extract, Load e Transform) possibilitam localizar e carregar os dados de forma bruta para que possam ser transformados depois. É um padrão mais utilizado em ambientes big data.

Com a adoção desses métodos, um fluxo de dados pode ser compreendido em 3 momentos:
1) Extração: atividade que inicia com a coleta das informações por meio de ferramentas de extração de dados que tenha particularidades do tipo *Web Crawling* (*web scraping* ou *screen scraping*) que se trata de uma técnica de extração de dados automatizada, geralmente, utilizada para coletar dados em sites. (Singrodia, Mitra, & Paul, 2019). Esse tipo de método tem a capacidade de colocar os dados extraídos em formato estruturado, incluindo, mas não limitado a XLSX, HTML e CSV. Esse momento é essencialmente para extrair e reunir conjuntos de dados da web que são um dos eixos para o *Big Data Analytics, Machine Learning* e Inteligência Artificial;
2) Armazenamento: essa etapa realizada em dois momentos: 1) assim que os arquivos são capturados e armazenados em um repositório onde ficam os dados brutos (*raw data*); 2) o processo de armazenamento continua quando é iniciado o processo de transformação, nesse momento pode-se usar uma estrutura de dados temporária conhecida como *Staging Area* ou Área de Teste onde os dados podem ser trabalhados;

3) Transformação: essa última etapa do processo, permeia desde o armazenamento, envolvendo ainda, se necessário, uma sequência de passos para o refinamento do dado de acordo com as regras de negócio e armazenados em uma estrutura definitiva.

Exposto o cenário dos dados abertos educacionais, adotou-se a implementação prática da metodologia de coleta dos dados que foi mencionada. O objetivo foi de adquirir os dados em ambiente web e apropriar esses dados em uma área comum para reuso da informação. O arcabouço computacional que envolve o processo de ELT irá seguir um conjunto de conceitos e abstrações para aplicação da metodologia exposta de forma a organizar os componentes e processos. Com esse entendimento, foram desenvolvidas tarefas padronizadas que realizaram a coleta dos dados abertos e disponibilizavam em um sistema de arquivos, posteriormente, armazenados em ambiente de banco de dados.

O levantamento realizado sobre as bases de dados permitiu identificar que os conjuntos de dados são disponibilizados em formatos variados com diferentes estruturas de informação e a plataforma onde são disponibilizados. Observando essas variâncias foi necessário implementar diferentes formas para extrair as informações dos arquivos.

Um dos cenários encontrados durante o processo de coleta. Foram identificados que algumas bases de dados são disponibilizadas em formato comprimido, no caso o ZIP. Então, houve a necessidade de desenvolver uma tarefa específica para realizar a atividade de descompactar esses tipos de arquivos para que a informação pudesse ser propriamente lida.

A etapa de coleta dos dados terminou com os conjuntos de dados extraídos e depositados em um sistema de arquivos. Para esse experimento foi utilizado o sistema de arquivos do próprio sistema operacional. Existem plataformas mais apropriadas com tecnologias mais robustas para desempenhar essa tarefa de concentração de dados heterogêneos. Segundo Almeida e Santos (2014), grande parte das implementações de data lake são baseadas na plataforma Hadoop (Plataforma de dados orientada a objetos altamente disponível). Os autores complementam que o Hadoop tem dois componentes principais - HDFS (Hadoop Distributed File System) e mecanismo MapReduce. Outras possibilidades de armazenamento são as tecnologias de banco de dados baseadas na abordagem NoSQL.

Com as bases coletadas, iniciou-se a segunda etapa do fluxo de dados. Foi implementada a tarefa para carregar os dados, lendo-os a partir do sistema de arquivos. O resultado dessa rotina é a importação dos dados abertos educacionais para um ambiente transitório de banco de dados aqui chamado de área de teste. Cabe observar que, nesse momento, podem ocorrer as primeiras transformações, por exemplo, a possibilidade de padronização dos atributos das bases (tipificação e nome) para que os dados possam ser trabalhados e refinados de acordo com a necessidade informacional da organização.

Por fim, os dados coletados foram armazenados em tabelas da área de teste que foram criadas no ambiente PostgreSQL, um sistema gerenciador de banco de dados objeto-relacional usado nesse experimento. No entendimento de Khine e Wang (2018), a ordem de processamento (Extrair-Carregar-Transformar) proporciona ao usuário a possibilidade de consultar os dados que serão transformados de acordo com as necessidades do usuário para o nível analítico. A partir desse momento os dados abertos educacionais estão disponíveis para serem trabalhados pelos especialistas de forma a enriquecer as bases de dados internas. Com as informações trabalhadas é indicado que os dados sejam então direcionados para outro ambiente de

armazenamento definitivo que pode ser um ambiente *Data Warehouse* ou bancos de dados NoSQL.

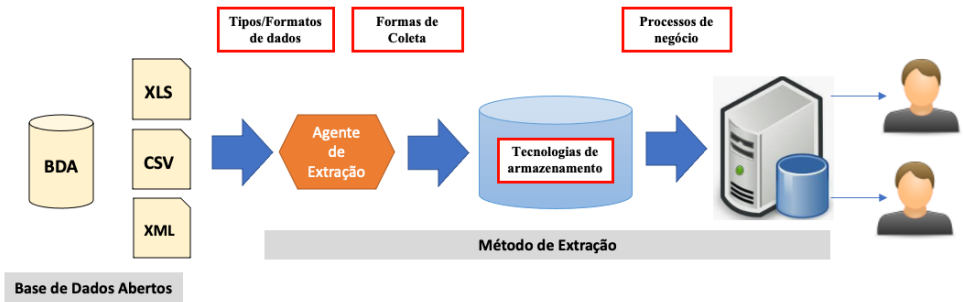
5. Framework Conceitual Preliminar

A construção do arcabouço se deu com base no entendimento de Regoniel (2016): 1) escolha do tema; 2) revisão da literatura; 3) isolamento dos elementos importantes; 4) pesquisa de campo; e por fim, 5) geração do framework conceitual. Conforme Swaen (2020), um dos passos nessa construção é a identificação das variáveis. McGahie, Bordage & Shea (2001: 923) estabelecem também a existência de relações entre as variáveis.

A Figura 3 apresenta o desenho do framework responsável pelo fluxo de consumo dos dados abertos educacionais desde a sua origem com as variáveis que foram percebidas até aqui. Este arcabouço preliminar se propõe a representar a forma como os arquivos de dados serão coletados, consumidos e armazenados. Nessa construção, foram identificados diversos elementos e variáveis.

5.1. Análise geral dos dados das entrevistas

A análise das entrevistas realizadas permitiu identificar alguns temas. Foram realizadas cinco entrevistas com especialistas com perfil em coleta, tratamento e disponibilização das informações, principalmente, nos produtos relacionados a dados abertos educacionais. Desta forma, o entrevistado A atua na sistematização e disseminação das informações até o produto ser oferecido em formato aberto, sendo essa a última etapa dos censos educacionais. O entrevistado B possui conhecimento em controle de qualidade e tratamento da informação atuando nas pesquisas educacionais (Educação Básica e Superior). Já o entrevistado C volta-se para o acompanhamento de todo o fluxo dos Exames e Avaliações da educação básica desde a preparação dos dados para aplicação dos certames, passando pelo cálculo das proficiências até a produção e divulgação dos microdados e sinopses estatísticas para a sociedade. Atuando sobre sistemas de informação e nas práticas tecnológicas, o entrevistado D viabiliza tecnologia necessária para manipulação e divulgação dos dados produzidos. Finalmente o entrevistado E atua no desenvolvimento de soluções focadas na consumação, análise e visualização dos dados coletados.



Elementos	
Base de Dados Abertos	as bases de dados abertos de interesse da área educacional.
Agente de Extração	automatização a nível de sistema
Método de Extração	método de extração-transformação-carga dos dados.

Variáveis	
Tipos e formatos dos dados	os diferentes tipos e formatos dos dados de desejo.
Formas de coleta	a forma como esses dados devem ser coletados.
Tecnologias de Armazenamento	a definição da tecnologia a ser adotada para o armazenamento das informações.
Processos de Negócio	definições de negócio.

Figura 3 - Desenho do framework preliminar

As observações dos entrevistados, foram analisadas, especificamente, as três questões iniciais possibilitaram um detalhamento das relações de entrada e saída que cercam os dados abertos produzidos pelo Inep, bem como a disposição tecnológica atual da Autarquia.

Pelo que foi observado quando perguntados sobre a interação com demais entes sobre o uso de dados abertos. A interação ocorre por meio de acordos com outras entidades. O entrevistado B destacou os canais de comunicação com entidades municipais e estaduais para realização dos Censos da Educação: "...para realizar a pesquisa a gente tem um processo de relacionamento intenso com as instituições de ensino na educação superior principalmente com uma relação mais direta e com as secretarias estaduais e municipais de educação..."

Os demais explicitaram que a iteração também ocorre por meio de um instrumento jurídico formalizado entre órgãos e entidades da Administração Pública. Já a sociedade civil acessa as informações a partir dos dados abertos, sinopses estatísticas e os indicadores educacionais.

O intuito de entender a interação com outras entidades para identificar a influência sobre boas práticas atualmente realizadas na publicação de dados. Nesse sentido, os entrevistados lembraram da iteração com o MCTI - Ministério da Ciência, Tecnologia e Inovações, entrevistado A: "porque nós estamos com o objetivo de organizar os nossos metadados...". Além do MCTI, foram lembrados, o IBGE, o Ministério da Economia e o IPEA. Acrescenta-se ainda a Lei Geral de Dados Abertos (LGDA) como fonte de boas práticas. Outra influência exemplificada em dados abertos vem dos órgãos de controle como TCU. Percebe-se que além das necessidades internas, a influência de outros órgãos públicos na aplicação de boas práticas vem principalmente por meio de recomendações e auditorias que permitem a evolução na prática de dados abertos.

Antes do prosseguimento de qualquer solução tecnológica é preciso determinar o papel da área de tecnologia da informação para entender cenários e contextos na qual ela está inserida. Segundo as declarações, a área de tecnologia é vista como um facilitador no provimento e desenvolvimento de soluções que permitem automatizar o trabalho sobre os dados abertos sendo definida pelos entrevistados A e C como fundamental. O entrevistado B destaca a área de TI no contexto do suporte de hardware e software por assim dizer: "...a tecnologia da informação ela tem um duplo papel tanto em relação a infraestrutura quanto em relação as próprias soluções para viabilizar a disseminação de dados abertos...".

Nesse contexto de qual é o papel da área de tecnologia, o entrevistado E a coloca como área que precisa ser posicionada dentro dos objetivos organizacionais: "..., mas hoje como é, teria toda infraestrutura pronta só precisava mesmo de ter uma equipe, um direcionamento, montar, fazer um planejamento e dedicar a esse projeto...". Desta maneira, a infraestrutura de TI assume a responsabilidade de prover tecnicamente a capacidade de trabalhar em ambientes distintos: um para os dados coletados e outro com dados que serão disponibilizados.

Desprende-se das considerações que o framework proposto atende às necessidades propostas no intuito de organizar o trabalho com os dados, não só com os dados abertos. Assim, os dados coletados vão de encontro a um dos objetivos dessa pesquisa que é o de promover a gestão da informação por meio de uma estrutura de big data e dados abertos: "...uma vez que eu tenho todos esses dados juntos eu acho que a TI colaboraria bastante também fornecendo ferramentas para que os usuários desse *big data* pudessem fazer seus cruzamentos, realizar suas consultas

de uma maneira fácil, de uma maneira simples mesmo para quem não tem tanto conhecimento em base de dados. Isso é algo que a gente sente muita falta....” (Entrevistado C)

5.2. Reconstrução do framework

Esse tópico reúne as observações dos entrevistados em relação ao framework conceitual preliminar com vistas a remodela-lo e chegar ao framework final. Para isto, foram analisadas as respostas do roteiro, especificamente, as três questões finais estão mais voltadas se o que está sendo proposto é válido para os especialistas e identificar possíveis melhorias. Na fase de desenvolvimento do fluxo de dados foram identificados elementos e variáveis que foram considerados componentes do framework conceitual proposto.

A realização do quarto e do quinto questionamentos tiveram o intuito de verificar se os elementos e variáveis dispostos na estrutura conceitual são aderentes a atividade a qual o fluxo de dados se propõe. No Quadro 6, observa-se que dois dos cinco entrevistados adotaram o pensamento que tanto os elementos quanto as variáveis são suficientes. Por outro lado, também foi o intuito dos questionamentos o de coletar junto aos especialistas, possíveis sugestões de melhoria na estrutura proposta.

Quadro 2 - Sínteses das respostas concedidas na 4º e 5º questões

Entrevistado	Sobre os Elementos	Sobre as Variáveis
A	*“...a princípio olhando aqui me parece que é esse o fluxo completo mesmo.”*	*“...à primeira vista eu também não consigo identificar nada que possa ter essa interferência, além disso, que você já mencionou...”*
D	*“...Nesse primeiro momento eles são suficientes tanto que dessa forma que a gente trabalha...”*	*“...não identifico não para a gente é suficiente.”*
Sugestões		
B	*“Então eu acho que tem que ter um elemento dentro do framework que faça o relacionamento da divulgação com as estratégias do órgão...”*	*“...Pensando que o processo de negócio é uma variável eu tô pensando que ela dá conta, por exemplo, de você avaliar se você está provendo soluções adequadas para necessidade dos usuários...”*
C	*“...deixar claro necessidade dessas ferramentas de consulta como se fosse uma ferramenta de BI...”*	*“Eu não sei se está dentro de processos de negócio, mas eu senti falta de metadados...Então acho fundamental que você tenha esses metadados como variável...”*
E	*"..., mas não sei se você pensou a questão da qualidade do dado, data quality porque o que a gente vê assim dado bruto principalmente da área educacional tem muita coisa que a gente faz esse tratamento na mão..."*	*“você vai tratar a questão da segurança desses dados agora com a questão da LGPD... A legal e de acesso também...”*

Fonte: elaborado pelos autores.

Em face do que foi dito, o intuito é ter um elemento que oriente a organização nas informações que sejam relevantes aos propósitos do Instituto e que se estabeleça em nível estratégico as demandas que possibilitem potencializar a consumação dos dados.

Outra necessidade apontada foi a necessidade de um elemento voltado a qualidade dos dados semelhante ao implementado na proposta de Benitez-Paez et. al (2018) e Tekinier e Keane (2013), em que foi estabelecido um componente com foco em metadados.

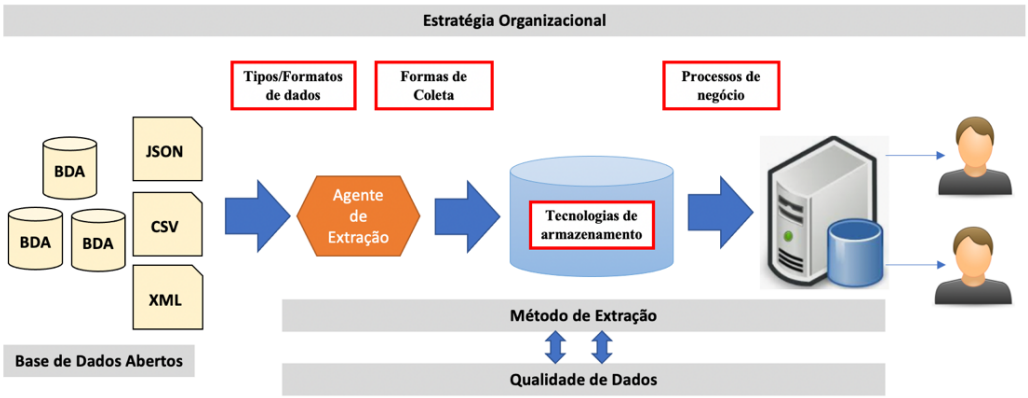
Como apresentado anteriormente, Imenda (2014) entende que o propósito de uma estrutura conceitual, sendo ela capaz de certa forma explicar ou prever, permitindo assim, o entendimento sobre o problema estudado. Nessa linha, o sexto questionamento possibilitou verificar a receptividade do framework frente ao entendimento que a representação significa sobre a tarefa de captura de dados. No entendimento de quatro entrevistados declararam que a representação é suficiente para realização da atividade.

Como já mencionado nessa pesquisa por Pasupuleti e Purra (2015: 10) e Machado (2017: 15-17), cabe esclarecer que um fluxo com características de *data lake* e *big data* podem ser estruturados por camadas, devendo ser delimitado a forma como serão usadas. Portanto, apesar de ser um objetivo desta pesquisa que o framework assume o papel que possibilite o enriquecimento dos dados abertos educacionais por meio de outros dados, a representação também pode servir apenas para coletar e disseminar dados abertos.

As percepções até aqui expostas demonstraram que dentro do contexto de cada especialista que foi entrevistado o framework conceitual atendeu no que diz respeito a atividade de captura de dados abertos e as considerações mencionadas foram a título de melhoria, alguns acréscimos para uma melhor adequação ao ambiente em que os dados abertos educacionais são coletados, preparados e disseminados.

5.3. Reorganização do framework conceitual preliminar

A reorganização do framework conceitual evidencia as sugestões fornecidas pelos entrevistados de acordo com as experiências e expertise de cada um. Para tanto, o framework conceitual foi reformulado, com base nas informações coletadas, e é apresentado na Figura 4.



Elementos	
Estratégia Organizacional	tratar dos objetivos de uso de dados de outros órgãos, direcionando a pesquisa e a disseminação de dados
Base de Dados Abertos Externas	as bases de dados abertos de interesse da área educacional.
Agente de Extração	automatização a nível de sistema
Método de Extração	método de extração-transformação-carga dos dados.
Qualidade de Dados	orientar toda a gestão de metadados e demais processos para "limpeza" e padronização dos dados.

Variáveis	
Tipos e formatos dos dados	os diferentes tipos e formatos dos dados de desejo.
Formas de coleta	a forma como esses dados devem ser coletados.
Tecnologias de Armazenamento	a definição da tecnologia a ser adotada para o armazenamento das informações.
Processos de Negócio	definições de negócio.

Figura 4 - Framework conceitual reorganizado após análise dos especialistas
Fonte: elaborado pelos autores.

A remodelagem permitiu o acréscimo de novos elementos e a melhor especificação sobre as variáveis. Portanto, os elementos: estratégia organizacional – irá tratar dos objetivos de uso de

dados de outros órgãos, direcionando a pesquisa e a disseminação de dados; as bases de dados abertos externas – as bases de dados abertos de interesse da área educacional; o agente de extração - automatização a nível de sistema para realizar a coleta das bases de dados; o método de extração - método de extração-transformação-carga dos dados; e a qualidade de dados – orientar toda a gestão de metadados e demais processos para "limpeza" e padronização dos dados.

Com relação às variáveis, a remodelagem foi em nível de entendimento sobre o escopo delas dentro do fluxo. Neste caso, tipos e formatos dos dados - os diferentes tipos e formatos dos dados disponíveis; as formas de coleta – os métodos como esses dados devem ser coletados; tecnologias de armazenamento – a escolha da tecnologia a ser adotada para o armazenamento das informações; e por fim os processos de negócio – que sobretudo devem cobrir as atividades relacionadas as definições de negócio sobre o uso dos dados.

6. Conclusões

O desenvolvimento de um arcabouço precisa de uma análise de cenário, identificando as possíveis entradas, saídas, componentes e fatores que influenciam no funcionamento da estrutura a ser desenvolvida. Neste ponto é importante a participação das partes interessadas: usuários e especialistas nas atividades e uso de dados abertos governamentais. A maioria das bases de dados abertos representam ou apoiam Políticas Públicas que devem ser continuamente acompanhadas de forma a medir seu grau de eficiência e eficácia. Fator importante a considerar em um cenário onde as boas práticas em se publicar dados abertos governamentais ainda são realizadas a níveis protocolares, uma das causas da existência de bases de dados abertos heterogêneas dispersas na web.

Os principais resultados foram a identificação do ambiente, da estrutura, dos elementos e das variáveis que envolvem a coleta e a disseminação de dados, que compõem a estrutura básica de um ambiente que recepcione dados estruturados e não estruturados para serem consumidos, sendo que são cinco elementos identificados: Estratégia Organizacional; Bases de Dados Abertos Externas; Agente de Extração; Método de Extração; e a Qualidade de Dados. Sendo quatro variáveis: Tipos e Formatos dos dados; Formas de Coleta; Tecnologias de Armazenamento; e Processos de Negócio.

Os Planos de Dados Abertos (PDA) dos entes educacionais selecionados reúnem informações valiosas sobre os dados abertos das entidades publicadoras. A partir deles também é possível obter detalhamento das bases de dados abertos governamentais. Por meio de um estudo sobre cada conjunto de dados, foi possível identificar e selecionar as bases de dados abertos de interesse da área educacional de cada ente educacional para o enriquecimento de informação.

A partir da análise da estrutura das bases de dados abertos governamentais foi possível observar o nível de conectividade de cada arquivo de dados. Seguindo os critérios estabelecidos pelo Esquema Cinco Estrelas, foi identificado que a maior parte dos conjuntos de dados foram classificados com nível 2 ou 3 estrelas, sendo esse último o mínimo para se considerar um dado aberto que permite conectividade. O fato de a maior parte das bases de dados não permitirem o uso efetivo do *linked data*, ressaltou-se a importância da criação de uma estrutura que permita a aplicação de tarefas voltadas a transformação desses dados para que eles "conversem" entre si antes do seu efetivo uso.

O desenvolvimento e a proposição de uma modelagem que possibilite a captura dos conjuntos de dados abertos de interesse foram por meio da reunião das informações coletadas durante a pesquisa ao se entender o contexto e as necessidades cabendo isolar os elementos que são importantes e também detectar as variáveis que influenciam no funcionamento dessa atividade. Assim, em um primeiro momento, possibilitando a proposição de uma estrutura conceitual preliminar.

As entrevistas com profissionais especializados foram esclarecedoras e permitiram identificar novos elementos e variáveis. Essa prática permitiu, de certa forma, aprofundar na atividade de coleta, preparação e disseminação de dados abertos o que possibilitou validar a proposta desse estudo o de desenvolver um framework conceitual.

O framework conceitual proposto pode assumir o papel que possibilite o enriquecimento dos dados abertos educacionais por meio da captura de outros dados sejam eles abertos ou de acesso limitado, mas a representação também pode servir apenas para coletar e disseminar dados abertos.

O cenário pesquisado e pelas características de um ambiente *big data*, o método ELT (extrair-carregar-transformar), diferente do ETL (extrair-transformar-carregar), mostrou-se mais eficiente para um cenário com variedade de tipos e formatos de bases de dados que possuem atributos heterogêneos que dependem de uma análise prévia para que possam ser transformados e agregados a outros dados. Dessa forma, houve melhora no desempenho da aquisição dos dados. Claro que uma vez identificados os dados de interesse, o objetivo é que essa atividade seja realizada uma única vez.

Do ponto de vista teórico, o framework conceitual desenvolvido serve para nortear como se deve implementar um fluxo de dados, preparando-o para atuar com bases de dados heterogêneas. Identificar o que você precisa para extrair essas informações. Do ponto de vista prático, a estrutura objetiva deixar os dados prontos para serem analisados pelos pesquisadores e demais usuários.

Por fim, o uso de dados abertos governamentais como insumo para enriquecimento de outros dados abertos serviu para fins acadêmicos mostrando a possibilidade de relação de informações, cabendo aos órgãos públicos realizarem acordos de cooperação para a consumação de dados mais detalhados que podem agregar ainda mais aos dados produzidos.

Este estudo foi construído voltado as entidades governamentais da área da educação sendo observadas apenas as bases de dados abertos governamentais de cada uma. A utilização desse framework conceitual é indicada, principalmente, para instituições que não possuem um processo de coleta e disseminação de dados definida e para aquelas que não usam as abordagens *big data* e *data lake*. Ele também apoia nas atividades de transformação e armazenamento. Dessa forma, sugere-se a realização de outros estudos afim de avaliar a aplicabilidade deste framework conceitual em outros órgãos do governo, que contemple não somente os órgãos do executivo da área da educação, mas também dos demais poderes da república.

Outra sugestão seria no aprofundamento de atividades voltadas ao gerenciamento de metadados que possam ser capturados e padronizados da mesma forma que os dados abertos. A estrutura aqui proposta apenas indica o elemento que irá tratar dos metadados não aprofundando nesse assunto. Cabe esclarecer que o *framework* conceitual proposto não chega ao nível analítico dos dados e com isso abre espaço para a proposição de práticas subsequentes voltadas a *big data*

analytics e *machine learning* para agregação de novas tecnologias que possibilitem a geração do conhecimento.

Referências Bibliográficas

Adeodato, P. J. L., Santos, F., Maílson, M., & Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: *XXV Simpósio Brasileiro de Informática na Educação*, 891-895.

Ahmed, E., Yaqoob, I., Hashem, I., Khan, I., Ahmed, A., Imran, M., et. al. (2017). The role of big data analytics in Internet of Things. *Computer Networks*, 129, 459-471.

Alcantara, W., Bandeira, J., Barbosa, A., Lima, A., Avila, T., Bittencourt, I., & Isotani, S. (2015). Desafios no uso de Dados Abertos Conectados na Educação Brasileira. In*: XXXV Congresso da Sociedade Brasileira de Computação*. 2015.

Almeida, F., & Santos, M. (2014). A conceptual framework for Big Data Analysis. In: Almeida, F., & Santos, M. Organizational, Legal, and Technological Dimensions of Information System Administration. IGI Global. Portugal, 199 – 223.

Attard, J., Orlandi, F., Scerri S., & Auer, S. (2015). A systematic review of open government data initiatives*. Government Information Quarterly*, 32 (4), 399-418.

Auer S. (2011). LOD 2 The Semantic Data Web. In: *Semantics & Media Conference* 2011. Disponível em: https://www.slideshare.net/lod2project/the-semantic-data-web-sren-auer-university-of-leipzig. Acesso em: 30 abr 2020.

Benitez-Paez, F., Comber, A., Trilles, S., & Huerta, J. (2018). Creating a conceptual framework to improve the re-usability of open geographic data in cities. *Transactions in GIS*, 22 (3), 806-822.

Berners-Lee, T. (2006). *Linked Data*. *W3C*. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 12 mar 2020.

Berends, J., Carrara, W., Vollers, H., Fechner, T., & Kleemann, M. (2017). *Analytical Report 5: Barriers in working with Open Data*. Luxembourg: Publications Office of the European Union.

BRASIL. Ministério do Planejamento Orçamento e Gestão. (2012). *Cartilha Técnica para Publicação de Dados Abertos no Brasil v1.0.* Brasília: Secretaria de Logística e Tecnologia da Informação.

Brasil. (2014). *Portaria nº 92*, de 24 de dezembro de 2014. Institui a arquitetura ePING (Padrões de Interoperabilidade de Governo Eletrônico).

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2015). *Censo da Educação Superior*. Portal do Inep. Brasília. Disponível em: http://portal.inep.gov.br/censo-da-educacao-superior. Acesso em: 28 ago 2020.

Brasil. (2016a). *Decreto nº 8.777,* de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo federal.

Brasil. Ministério do Planejamento, Orçamento e Gestão. (2016b). *VCGE - Vocabulário Controlado do Governo Eletrônico*. Brasília: Ministério do Planejamento, Orçamento e Gestão.

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Capes. (2017). *Plano de Dados Abertos.* Brasília: Capes.

Brasil. Ministério do Planejamento, Desenvolvimento e Gestão. (2018a). *Padrões de Interoperabilidade de Governo Eletrônico – ePING - Documento de Referência.* Brasília: Ministério do Planejamento, Desenvolvimento e Gestão.

Brasil. Fundo Nacional de Desenvolvimento da Educação. (2018b). *Plano de Dados Abertos do Fnde 2018/2019.* Brasília: Fundo Nacional de Desenvolvimento da Educação.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2020a). *Política e Plano de Dados Abertos do Inep (Biênio – 2020-2021).* Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Brasil. Ministério da Educação. (2020b). *Plano de Dados Abertos 2020 a 2022*. Brasília: Ministério da Educação. Disponível em: https://www.gov.br/mec/pt-br/media/acesso_informacacao/pdf/PDA_MEC_2020_2022_ED.pdf. Acesso em: 28 abr 2020.

Brasil. Secretária de Transparência e Prevenção da Corrupção. (2020c). *Manual de elaboração de Planos de Dados Abertos (PDAs)*. 2020c. Disponível em: https://www.gov.br/cgu/pt-br/centrais-de-conteudo/publicacoes/transparencia-publica/arquivos/manual-pda.pdf. Acesso em: 24 jul 2020.

Brasil. Controladoria-Geral da União. Portal Brasileiro de Dados Abertos. (2020d). *O que é INDA?* Disponível em: http://wiki.dados.gov.br/#:~:text=O%20que%20%C3%A9%20a%20INDA,o%20disposto%20na%20e%2DPING%20. Acesso em: 02 de ago 2020.

Brasil. Infraestrutura Nacional de Dados Espaciais. (2020e). *A INDE. Apresentação. O Portal Brasileiro De Dados Geoespaciais - Sig Brasil*. Brasília. Apresentação. Disponível em: https://inde.gov.br/Inde/Apresentacao. Acesso em: 02 ago 2020.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2020f). *Censo Escolar*. Portal do Inep. Brasília.

Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). (2020g). *Indicadores Educacionais.* Portal do Inep.

Casaes, J. C. C. (2019). *Governança De Dados Abertos Governamentais: Framework Conceitual Para As Universidades Federais, Baseado Em Uma Visão Sistêmica.* Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil.

Chudrik, A. (2013). *Planejamento e Aplicação de Recursos do Programa Dinheiro Direto na Escola em uma Instituição Municipal de Tempo Integral em Telêmaco Borba*. Trabalho de Conclusão de Curso (Especialização) – Universidade Tecnológica Federal do Paraná, Curitiba, PR, Brasil.

Dados Abertos. (2017). *O que são dados abertos?* Portal Brasileiro de Dados Abertos. Disponível em: http://www.dados.gov.br/pagina/dados-abertos. Acesso em: 12 mar 2020.

Dumbill E. (2014). *Understanding the Data Value Chain*. IBM Big Data & Analytics Hub. Disponível em: https://www.ibmbigdatahub.com/blog/understanding-data-value-chain. Acesso em: 05 abr 2020.

Ferreira, R. P. O. (2016). *Educação infantil pública e privada na RMBH: uma análise a partir do Censo Escolar 2014.* Dissertação de Mestrado, Escola de Governo Professor Paulo Neves de Carvalho, Fundação João Pinheiro, Belo Horizonte, MG, Brasil.

Gartner. (2020). Data Lake. In: Gartner Glossary. Gartner. Disponível em: https://www.gartner.com/en/information-technology/glossary/data-lake. Acesso em: 15 ago 2020.

Geothink. (2016). *What is Open Data and Open Licensing?* Citizen's Guide to Open Data. Disponível em: https://citizens-guide-open-data.github.io/guide/1-open-data. Acesso em: 16 ago 2020.

Gil, A. C. (2002). *Como elaborar projetos de pesquisa*. 4. ed. São Paulo: Atlas.

Hitzler, P, & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, 4, 233–235.

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2016). *Big Data Para Leigos*. Rio de Janeiro: Alta Books.

Imenda, S. (2014). Is There a Conceptual Difference between Theoretical and Conceptual Frameworks? *Journal of Social Sciences*, 38 (2), 185-195.

Janssen, M., Konopnicki, D., Snowdon, J. L., & Ojo, A. (2017). Driving public sector innovation using big and open linked data (BOLD). *Information Systems Frontiers*, 19, 189-195.

Janssen, M., & Kuk, G. (2016). Big and Open Linked Data (BOLD) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce*, 26 (1-2), 3-13.

Janssen, M., & Hoven, J. V. H. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly*, 32, 363-368.

Johson, R. E. (1997). Components, Frameworks, Patterns. *Communications of the ACM*, 40, 10-17.

Khine, P. P., Wang, Z. S. (2018). Data Lake: A New Ideology in Big Data Era. *In: 4th Annual International Conference on Wireless Communication and Sensor Network (WCSN 2017)*, 17, 03025.

Machado, A. L. (2017). *Administração do Big Data*. São Paulo: Senac.

McGaghie, W.C., Bordage, G., & Shea, J. A. (2001) Problem statement, conceptual framework, and research question. **Academic Medicine**. v. 76. p. 923–924, 2001.

Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, 88, 300-305.

Mockus M., Palmirani, M. (2015). Open Government Data Licensing Framework. *In:* Kő, A., Francesconi, E. *Electronic Government and the Information Systems Perspective***.** EGOVIS 2015. Lecture Notes in Computer Science, 9265.

Muniz, V. M., & Carvalho, A. T. (2007). O Programa Nacional de Alimentação Escolar em município do estado da Paraíba: um estudo sob o olhar dos beneficiários do Programa. *Revista de Nutrição*, 20 (3), 286-296.

Oliveira P. (2016). Babel não revisitada. *Revista Graphos*, 18 (2), 24-42.

Pasupuleti, P., & Purra, B.S. (2015). Data Lake Development with Big Data. Packt Publishing.

Rautenberg, S., Souza, L., Dall'agnol, J .M. H, & Michelon, G. A. (2018). *Guia Prático para Publicação de Dados Abertos Conectados na Web.* Curitiba: Appris.

Regoniel, P. (2016). *Conceptual Framework Development Handbook – A Step by Step Guide with Practical Examples.* Simplyeducate.me.

Sansone, S.., Rocca-Serra, P., Field, D., *et al.* (2012). Toward interoperable bioscience data. *Nature Genetics*, 44 (2), p. 121–126.

Santos, P., Ferreira, R., & Miranda, P. (2017). Dados Abertos Educacionais: Uma Revisão da Literatura Brasileira. *In: XXVIII* ***Simpósio Brasileiro de Informática na Educação***, 11-20.

Science International. (2015). *Open Data in a Big Data World.* 2015. Disponível em: https://council.science/wp-content/uploads/2017/04/open-data-in-big-data-world_long.pdf. Acesso em: 18 abr 2020.

Silva, G. A. (2015). *O Programa Dinheiro Direto na Escola (PDDE) como mecanismo da descentralização financeira, participação e autonomia na gestão escolar.* Dissertação de Mestrado, Universidade Federal de Alagoas, Maceió, AL, Brasil.

Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications. In: *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 1-6.

Swaen, B. (2020). *Conceptual framework*. Scribbr. Disponível em: https://www.scribbr.com/dissertation/conceptual-framework/. Acesso em: 06 jul 2020.

Tekinier, F., & Keane, J. A. (2013). Big Data Framework. In: *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 1494-1499.

Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, 22, OECD Publishing.

Varpio, L., Paradis, E., Uijtdehaage, S., & Young, M. (2020). The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. *Academic Medicine*. 95 (7), 989-994.

W3C Brasil. (2011a). *Manual dos dados abertos: governo*. Cooperação Técnica – NIC.br. São Paulo: Comitê Gestor da Internet no Brasil.

W3C Brasil. (2011b). *Manual dos dados abertos: desenvolvedores*. Cooperação Técnica – NIC.br. São Paulo: Comitê Gestor da Internet no Brasil.

Werle, F. O. C. (2014), Panorama das políticas públicas na educação brasileira: uma análise das avaliações externas de sistemas de ensino. *Revista Lusófona de Educação*, v. 27, 159-179.

Williamson, B. (2015). Governing software: Networks, databases and algorithmic power in the digital governance of public education. *Learning, Media and Technology*, 40 (1), 83-105.

Zaidan, F. H, & Bax M. P. (2013). Linked Open Data como forma de agregar valor às informações clínicas. *ATOZ: Novas Práticas em Informação e Conhecimento***.** 2 (1), 44-59.