UNIVERSIDADE FEDERAL DE MATO GROSSO – UFMT
FACULDADE DE ADMINISTRAÇÃO E CIÊNCIAS CONTÁBEIS – FACC
PROGRAMA DE PÓS-GRADUAÇÃO EM PROPRIEDADE INTELECTUAL E TRANSFERÊNCIA DE TECNOLOGIA PARA A INOVAÇÃO – PROFNIT

Fabio Antonio Rodrigues

Um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

Cuiabá
2022

Fabio Antonio Rodrigues

Um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

Dissertação apresentada ao Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação - Ponto Focal Cuiabá, na Universidade Federal de Mato Grosso, para obtenção do Grau de Mestre em Propriedade Intelectual e Transferência de Tecnologia para a Inovação.

Orientador: Prof. Cristiano Maciel, Dr.

Cuiabá

2022

Dados Internacionais de Catalogação na Fonte.

R696m Rodrigues, Fabio Antonio.
Um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT / Fabio Antonio Rodrigues. -- 2022
76 f. ; 30 cm.

Orientador: Cristiano Maciel.
Dissertação (mestrado profissional) – Universidade Federal de Mato Grosso, Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação, Cuiabá, 2022.
Inclui bibliografia.

1. Dados Abertos Governamentais. 2. Transparência. 3. Tecnologias da Informação e da Comunicação. 4. ETL. I. Título.

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL DE MATO GROSSO

PRÓ-REITORIA DE ENSINO DE PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM [NOME DO PPG]

FOLHA DE APROVAÇÃO

TÍTULO: Um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

AUTOR: MESTRANDO Fábio Antonio Rodrigues

Dissertação defendida e aprovada em **11 de março de 2022**.

COMPOSIÇÃO DA BANCA EXAMINADORA

- **Prof. Dr. Cristiano Maciel** - Orientador e Presidente da Banca

Instituição: Universidade Federal de Mato Grosso

- **Prof. Dr. Paulo Augusto Ramalho de Souza** - Membro Interno

Instituição: Universidade Federal de Mato Grosso

- **Prof. Dr. Juliano Fisher Naves** - Membro Externo ao Ponto Focal Cuiabá e Membro da Rede PROFNIT

Instituição: Instituto Federal de Rondônia

- **Prof. Dr. José Viterbo Filho** - Membro Externo

Instituição: Universidade Federal Fluminense

Cuiabá, 11 de março de 2022.

Documento assinado eletronicamente por **José Viterbo Filho**, **Usuário Externo**, em 23/03/2022, às 12:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **PAULO AUGUSTO RAMALHO DE SOUZA**, **Docente da Universidade Federal de Mato Grosso**, em 23/03/2022, às 14:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Juliano Fischer Naves**, **Usuário Externo**, em 23/03/2022, às 16:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **FABIO ANTONIO RODRIGUES**, **Técnico Administrativo em Educação da CES / STI / REITORIA - UFMT**, em 25/03/2022, às 10:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **CRISTIANO MACIEL**, **Docente da Universidade Federal de Mato Grosso**, em 25/03/2022, às 11:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufmt.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4527415** e o código CRC **E447A4A2**.

Referência: Processo nº 23108.015663/2022-40 SEI nº 4527415

Reitor da Universidade Federal de Mato Grosso

Prof. Dr. Evandro Aparecido Soares da Silva

Pró-reitor (a) de Pós-graduação e Pesquisa

Prof. Dr. Jackson Antonio Lamounier Camargos Resende

Gestora do Programa de Pós-Graduação em Propriedade Intelectual

e Transferência de Tecnologia para a Inovação – PROFNIT (Ponto focal Cuiabá-MT)

Prof. Dra. Luciane Cleonice Durante

Fabio Antonio Rodrigues

RESUMO

Existe um grande volume de dados sendo gerados a cada momento na rede mundial de computadores no mundo globalizado. Entre esses dados, temos os dados abertos que são os que podem ser usados e distribuídos livremente pela sociedade. As informações de domínio público originadas no âmbito governamental são denominadas dados abertos governamentais. A disponibilização dos dados abertos governamentais contribui para o aumento da transparência do estado e a participação e controle social sobre as instituições públicas, sendo de extrema importância para a democracia moderna. Os recursos das Tecnologias da Informação e da Comunicação juntamente com a internet têm contribuído para um acesso mais rápido e integral às informações públicas do que a mídia convencional tanto a nível local, nacional ou internacionalmente. Diante deste cenário, a Universidade Federal de Mato Grosso como uma instituição pública possui a responsabilidade de criar ações para a implementação e promoção da abertura de seus dados conforme às leis nacionais, porém não foi identificado a existência de modelos ou métodos automatizados para geração de dados abertos governamentais no contexto institucional. Neste sentido, o presente trabalho apresenta uma metodologia para geração de dados abertos governamentais conforme um metaprocesso e seguindo uma abordagem tecnológica chamada de Extração, Transformação e Carga de dados com o objetivo de gerar dados abertos governamentais no contexto dos sistemas da UFMT. Do ponto de vista metodológico, utilizou-se a pesquisa bibliográfica como método de coleta de dados, tendo como referencial ferramentas, modelos, processos ou métodos para geração de dados abertos governamentais, respaldando a fundamentação teórica da solução. O método proposto apresenta uma solução genérica para o enriquecimento semântico dos dados extraídos, com a aplicação desenvolvida decorrente do método gerando dados abertos governamentais de qualidade prontos para publicação em um catálogo, utilizando os recursos da ferramenta de código aberto Kettle *Data Integration* da Pentaho. O método foi verificado e considerado adequado por um grupo focal composto por diversos especialistas. Ressaltando a contribuição do trabalho, a demonstração do método via aplicação atestou a viabilidade técnica da solução com a disponibilização de dados em formato aberto, estruturado e compreensível por agentes de *software* ou humanos ao final do processo automatizado.

Palavras-chave: Dados Abertos Governamentais. Democracia. Transparência. Tecnologias da Informação e da Comunicação. Kettle.

ABSTRACT

There is a large volume of data being generated every moment on the world wide web in the globalized world. Among this data, we have open data which is what can be used and distributed freely by society. Public domain information originated within the scope government data is called open government data. The availability of open data government contributes to increasing the transparency of the state and social participation and control about public institutions, being extremely important for modern democracy. The resources of Information and Communication Technologies together with the internet have contributed to a faster and more comprehensive access to public information than conventional media at both the local level, nationally or internationally. Given this scenario, the Federal University of Mato Grosso as a public institution has the responsibility to create actions for the implementation and promotion of opening of your data in accordance with national laws, but the existence of models was not identified or automated methods for generating open government data in the institutional context. In this sense, this paper presents a methodology for generating open data according to a metaprocess and following a technological approach called Data Extraction, Transformation and Loading with the objective of generating open government data in context of UFMT systems. From a methodological point of view, bibliographic research was used as a method of data collection, having as theoretical framework: tools, models, processes or methods for generating open government data, supporting the theoretical foundation of the method proposed by the author. The proposed method presents a generic solution for enrichment semantic data extracted, and resulting from the method, the application developed generate quality government open data ready for publication in a catalog, using the features of the Pentaho's open source Kettle Data Integration. The method was verified and considered adequate by a focus group composed of several specialists. As a contribution to the work, the case study demonstrated the technical feasibility of the method with the availability of data in an open, structured and understandable format by software agents or humans at the end of the automated process.

Keywords: Open Government Data. Democracy.Transparency. Information and Communication Technologies. Kettle.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API *Application Programming Interface* (Interface de Programação de Aplicações)

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CKAN *Comprehensive Knowledge Archive Network*

CSV Character-separated values (valores separados por um delimitador)

DAC Dados Abertos Conectados

DAE Dados Abertos Educacionais

DAG Dados Abertos Governamentais

ETL Extração, Transformação e Carga (Extract, Transform e Load)

HTTP *Hypertext Transfer Protocol*

IBGE Instituto Brasileiro de Geografia e Estatística

INDA Infraestrutura Nacional de Dados Abertos

INEP Instituto Nacional de Estudos e Pesquisas Educacionais

JSON JavaScript *Object Notation*

LAI Lei de Acesso à Informação

LOD *Linked Open Data*

OUAI *Ontology for Universities and Academic Information*

PDA Plano de Dados Abertos

PDF Portable Document Format

PING Padrões de Interoperabilidade de Governo Eletrônico

PLN Processamento de Linguagem Natural

RDF *Resource Description Framework*

TCU Tribunal de Contas da União

TIC Tecnologia da Informação e Comunicação

UAI Unidade Acadêmica de Informática

UFMT Universidade Federal de Mato Grosso

UnB Universidade de Brasília

URI *Uniform Resource Identifier*

URL *Uniform Resource Locator*

XML *Extensible Markup Language*

W3C *World Wide Web Consortium*

SUMÁRIO

1 INTRODUÇÃO

A globalização, a alta conectividade e a popularização da internet não só facilitaram a criação, publicação e compartilhamento de informações como produziram um grande volume de dados disponibilizados na rede mundial de computadores (SILVA *et al*., 2016).

Em meio a essa quantidade maciça de dados disponíveis, há os emitidos pelo governo, os quais, conforme sugerem Dutra e Lopes (2013), devem ser disponibilizados para a sociedade por meio de canais de comunicação e ferramentas interativas com o cidadão. Com o uso das Tecnologias da Informação e da Comunicação (TICs) é possível promover meios para a participação popular nas questões governamentais, de modo a ajudar o cidadão na tomada de decisões de seu interesse e a sociedade no rumo para a conquista de uma democracia real (MACIEL, 2008).

Considerando que é papel do Estado construir uma relação mais próxima com os cidadãos, faz-se necessário propor mecanismos para ampliação da transparência e acesso aos dados públicos, baseados em tecnologia de perspectiva inovadora, que visem a facilitar o acesso aos dados e contribuir significativamente com a aproximação da sociedade ao governo. Com o avanço tecnológico e com as oportunidades criadas pela internet é possível produzir a democratização da informação, tendo em vista que "as novas plataformas tecnológicas, sua expansão, redução de custo e facilidade de acesso contribuíram para o desenvolvimento de um novo modelo de sociedade baseado na informação e no conhecimento" (BERBERIAN *et al*., 2014).

Desse contexto surge o conceito de dados abertos, assim definido pela *Open Knowledge* Brasil (2020, n.p.): "[...] dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa e sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras". Denominados de "dados abertos governamentais" (DAG) no âmbito governamental, eles contribuem para o aumento da transparência dos governos e para a participação e controle social sobre as instituições públicas. No ambiente educacional os dados públicos são chamados de "dados abertos educacionais" (DAE), colaborando com a inovação tecnológica na educação.

Em uma Instituição de Ensino Superior (IES), existem várias informações que podem ser vistas como dados abertos governamentais e disponibilizadas de forma aberta (currículos de professores, ementas de disciplinas, catálogo dos cursos de graduação, projetos etc.), uma vez que, além de reutilizáveis pela própria instituição ou por outras, favorecem a criação de canais que ajudam a sociedade a conhecer melhor a instituição, suas atividades e informações importantes (ALENCAR *et al*., 2018).

A Universidade Federal do Mato Grosso (UFMT), na condição de IES pública, possui a responsabilidade de criar ações para a implementação e promoção da abertura de seus dados em obediência às leis nacionais, a exemplo da Lei de Acesso à Informação (BRASIL, 2011). Motivado por essa exigência legal, o presente trabalho propõe um método para captura e compartilhamento de dados abertos no contexto da UFMT com o objetivo de deixá-los ao alcance do público em geral.

1.1 Contextualização e problema de pesquisa

Segundo Araújo e Souza (2011, p. 2), "as TICs promoveram uma revolução nos meios de informação, construindo uma nova relação entre governo e cidadãos. Esta nova relação deu origem ao chamado Governo Eletrônico, que possibilita uma administração pública mais acessível, eficiente, democrática e transparente". Mais do que a mídia convencional, as TICs e a internet têm contribuído para o acesso mais rápido e integral às informações públicas em termos de abrangência (local, nacional e internacional).

No Brasil o acesso aos dados governamentais é um direito garantido pela Lei de Acesso à Informação (BRASIL, 2011), que alterou o paradigma do Estado na posição de dono das informações e tornou regra a cultura do acesso e fez aumentar a demanda por procedimentos para desburocratizar e garantir o acesso às informações de natureza pública (GONÇALVES; GAMA, 2018). Sendo a tecnologia um caminho para sociedades mais democráticas, nas quais a informação flui mais livremente, possibilitando que seus cidadãos tomem decisões políticas mais bem informados.

Disponibilizados gratuitamente por entidades públicas e suscetíveis de serem utilizados, reutilizados e redistribuídos livremente pelos cidadãos (KLEIN et al., 2018), os DAGs se prestam ao provimento da transparência dos dados do governo.

Apesar da importância ao livre acesso à informação, estudos apontam a existência de uma transparência incompleta e desigual entre as esferas de governo e ainda voltada para atender as exigências da lei (COELHO *et al*., 2018). E semelhante constatação se verifica em Maciel (2008, p. 5): "O diagnóstico deixa clara a carência de soluções eficientes, efetivamente disponíveis e inovadoras para alavancar a participação popular[...]". Tomando as evidências dos estudos mencionados, pretende o presente trabalho suprir as lacunas da transparência incompleta e da falta de processos automatizados por meio de um método replicável para prover o acesso a informações educacionais no contexto da UFMT.

Consoante a definição de dados abertos governamentais apresentada no portal do governo digital - "uma metodologia para a publicação de dados do governo em formatos reutilizáveis, visando

o aumento da transparência e maior participação política por parte do cidadão" (GOV.BR, 2021, n.p.)-, podemos considerar que a UFMT não dispõe de uma metodologia automatizada para a geração de Dados Abertos Educacionais (DAE), sendo que a instituição possui um grande volume de informações dispersas no seu portal, sistemas e banco de dados acadêmico com potencial para publicação.

Filtrar, padronizar e disponibilizar parte dessa informação constitui um problema para as instituições públicas, pois uma boa parcela dela está armazenada nos bancos de dados (não estando disponíveis para consultas diretas) junto com informações confidenciais. Tais dificuldades suscitaram então a pergunta principal deste trabalho: como produzir e disponibilizar de forma automatizada dados abertos educacionais na UFMT?

A UFMT possui um catálogo de dados abertos disponível para consultas públicas, porém a manutenção do conteúdo é feita manualmente pelas unidades. A instituição orienta que sejam usados o vocabulário e os metadados do governo eletrônico, mas não possui uma ontologia ou modelos próprios. Para solucionar a questão, este trabalho sugere um método para captura e compartilhamento de dados abertos educacionais no contexto dos sistemas da UFMT e a automatização via aplicação ETL (*Extract - Transform - Load*). O processo automatizado gera ganhos de eficiência em comparação a extrações pontuais.

1.2 Objetivos da pesquisa

O objetivo principal deste trabalho é propor um método de coleta, tratamento e publicação dos dados abertos para a integração de diversos sistemas e bases de dados no contexto da UFMT. Em paralelo, pretende-se dar condições à criação de um cenário para captura e compartilhamento de dados abertos educacionais por meio de um ambiente tecnológico com proposta de abordagem baseada no processo de ETL.

De forma geral, os objetivos específicos são:

- Mapear e definir o processo ETL para extração e conversão (enriquecimento) dos dados;
- Criar uma aplicação ETL *desktop* para extração e enriquecimento de uma amostra de dados (indicadores educacionais), visando à publicação dos indicadores gerados em uma ferramenta de catalogação de dados.
- Validar o método por análise estatística das respostas, oriundas de um questionário virtual de pesquisa de opinião na escala Likert, aplicado a um grupo focal.

1.3 Justificativa

Um ambiente projetado para a publicação e consumo de dados abertos gera um ganho na transparência e colaboração com a sociedade, considerando que o acesso à informação é um dos pilares da democracia moderna (PENTEADO, 2020; FERREIRA, 2017).

Com a utilização de recursos tecnológicos, é possível a automatização da captura e modelagem dos dados para a divulgação no portal da UFMT, atendendo a Lei nº 12.527/2011, conhecida como Lei de Acesso à Informação (BRASIL, 2011), e o consequente aumento da transparência dos dados da instituição.

Modelar dados de um domínio ou contexto em formato conectado, com o uso de metadados e padronizações, viabiliza a produção de dados abertos governamentais conectados com qualidade assegurada, aumentando as chances de seu consumo por humanos ou computadores, o que vai ao encontro do objetivo do Plano de Dados Abertos da UFMT, cujo propósito é "promover a abertura de dados da Universidade Federal de Mato Grosso, zelando pelos princípios da publicidade, transparência e eficiência, visando o aumento da disseminação de dados e informações para a sociedade, bem como a melhoria da qualidade dos dados disponibilizado" (UFMT, 2021, n.p.).

A inovação reside na criação do método e na da aplicação ETL, que minera as bases de dados com o objetivo de extrair as informações públicas que vão alimentar a plataforma de dados abertos governamentais da UFMT e, dessa forma, gerar dados abertos educacionais de qualidade. A importância do presente estudo está no seu contributo para a superação das dificuldades técnicas e de infraestrutura na abertura dos DAG na Universidade Federal de Mato Grosso.

1.4 Estrutura do Trabalho

A estrutura geral do trabalho é composta pelas seguintes seções:

- Introdução**:** onde encontram-se descritos os elementos que contribuem para a visão geral da dissertação: contextualização e problemática, objetivos, justificativas e estrutura do trabalho.
- Revisão de literatura**:** apresenta a fundamentação teórica e os principais conhecimentos utilizados na pesquisa, assim como os exemplos da aplicabilidade dos conceitos sobre dados abertos no cenário nacional e os trabalhos relacionados ao tema da dissertação.
- Procedimentos metodológicos: apresenta os procedimentos metodológicos adotados com o auxílio de um fluxograma.
- Proposta do método: apresenta o método proposto em detalhes, com ilustrações e diagramas do que é esperado em cada fase.

- Análise dos resultados: exibe os resultados obtidos com o método e sua aplicação, analisando a implementação e as saídas de cada fase. Os resultados da oficina com o grupo focal também são comentados neste capítulo.
- Considerações finais: discorre sobre as considerações finais, enfatiza o objetivo e as contribuições da pesquisa e do método apresentado e ainda comenta as limitações do trabalho.

Por fim, as referências que subsidiaram o embasamento teórico desta pesquisa e os apêndices que complementam o trabalho.

2 REVISÃO DE LITERATURA

A visão baseada no conhecimento circulando livremente pela internet traz consigo mudanças culturais no âmbito do acesso à informação pela sociedade. A grande quantidade de informação gerada na rede de computadores na contemporaneidade mantém os cidadãos mais informados do que as gerações anteriores, contudo, para maximizar o potencial de transformação dos dados é preciso refletir sobre como as informações estão sendo publicadas.

Obter informação de qualidade na atualidade é simultaneamente importante e desafiador em meio a tantas publicações disponíveis na internet. Assim, há que se concordar com Jamil e Neves (2000, p. 45) que "chegou a vez da informação":

> Deve-se avaliar o seu poder no processo de tomada de decisões, como ela é gerada, formatada, processada, armazenada e oferecida ao grande público, além de se avaliar como isto afeta a vida do indivíduo, tanto como consumidor, eleitor, contribuinte, quanto como agente de decisão dentro de seu grupo

Pensando em práticas para promover o conhecimento livre e a difusão da informação, temos organizações como a *Open Knowledge* Brasil (2020, n.p.) atuando em vários países e com a seguinte visão: "Querer um mundo onde o conhecimento livre esteja presente em todo nosso cotidiano. Promovemos o conhecimento livre por acreditar em sua capacidade de gerar grandes benefícios sociais". É oportuno dizer que tal afirmação contribui para a perspectiva adotada na abordagem do tema dados abertos no presente estudo.

Dando continuidade ao que prega a *Open Knowledge* Brasil (2020, n.p.), vejamos como ela relaciona o termo conhecimento "aberto" com as seguintes ideias:

> - Conhecimento é um bem-comum, ou seja, qualquer pessoa pode usar e participar de sua construção;
> - Informatizados ou não, os sistemas devem ser "interoperáveis", o que significa ampliar ao máximo sua capacidade de se comunicar de forma transparente e de se conectar com outros sistemas.

Adicionalmente, ela apresenta a seguinte definição de dados abertos: "qualquer pessoa está livre para acessá-lo, utilizá-lo, modificá-lo e compartilhá-lo (restrito, no máximo, a medidas que preservam a proveniência e abertura)" (OPEN KNOWLEDGE BRASIL, 2020, n.p.).

Partindo destes conceitos, constata-se que as definições encontradas na página da *Open Knowledge* Brasil (2020, n.p.) apontam características como disponibilidade de acesso, reutilização e redistribuição dos dados e participação universal, sendo:

> - **Disponibilidade de acesso**: Os dados devem estar disponíveis como um todo e sob custo não maior que um custo razoável de reprodução, preferencialmente possíveis de serem baixados pela internet. Os dados devem também estar disponíveis de uma forma conveniente e modificável;

- **Reutilização e redistribuição**: Os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, inclusive a combinação com outros conjuntos de dados.
- **Participação Universal**: Todos devem ser capazes de usar, reutilizar e redistribuir. Não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições de uso 'não-comercial' que impediriam o uso 'comercial', ou restrições de uso para certos fins (ex.: somente educativos) excluem determinados dados do conceito de "abertos".

À análise das características, se acrescenta que os dados são classificados conforme as três leis dos dados abertos (que são testes ou validações e não leis no sentido de promulgadas pelo Estado) para considerar um dado como aberto ou não. Elas foram propostas por David Eaves, especialista em políticas públicas e ativista dos dados abertos (DADOS.GOV.BR, 2021, n.p.). O Portal Brasileiro de Dados Abertos (2021, n.p.) assim as enumera:

- Se o dado não pode ser encontrado e indexado na *web*, ele não existe;
- Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; e
- Se algum dispositivo legal não permitir sua replicação, ele não é útil.

Os dados públicos organizados e disponíveis na internet em formato aberto e livres para acesso e compartilhamento criam um canal com a população, aproximando a sociedade do governo, trazendo benefícios sociais como transparência, participação da população nas decisões públicas e a tomada de decisão fundamentada na informação. Na esfera pública, dados abertos governamentais "são os produzidos, coletados ou custodiados por autoridades públicas e disponibilizados em formato aberto" (TCU, 2015, n.p.).

Quanto às características dos DAGs publicados, resultante dos oito princípios definidos por um grupo de trabalho, em 2007, nos EUA, temos as abaixo descritas:

- **Completos**. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo, mas não se limitando a, documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, reguladas por estatutos.
- **Primários**. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.
- **Atuais**. Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.
- **Acessíveis**. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.
- **Processáveis por máquina**. Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.
- **Acesso não discriminatório**. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.
- **Formatos não proprietários**. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.
- **Licenças livres**. Os dados não estão sujeitos a restrições por regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de

privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos (DADOS.GOV.BR, 2021, n.p.).

E por que abrir os dados públicos? O Tribunal de Contas da União (TCU, 2015, n.p) apresenta "5 motivos para a abertura de dados na administração pública". Os motivos são:

- Transparência na gestão pública;
- Contribuição da sociedade com serviços inovadores ao cidadão;
- Aprimoramento na qualidade dos dados governamentais;
- Viabilização de novos negócios;
- Obrigatoriedade por lei.

Abrir os dados governamentais é de extrema importância para a democracia no mundo globalizado. A disponibilização de dados abertos apresenta diversos benefícios, como a "melhoria da gestão pública, o provimento da transparência, o estímulo ao controle e participação social, a geração de emprego e renda e o fomento à inovação tecnológica" (TCU, 2015, n.p.).

Um dos aspectos importantes da disseminação da informação é o conhecimento chegando até o cidadão. E com o avanço da tecnologia surgem ferramentas para coleta, armazenamento e processamento de dados em grandes volumes, velocidade e com potencial de geração de conhecimento para a tomada de decisão, indo ao encontro do que a sociedade exige em relação à transparência na gestão pública (TCU, 2015). Com a contribuição das TICs é possível uma maior aproximação entre o Estado e a sociedade, facilitando serviços, acelerando processos e aumentando a transparência e a participação social (BERBERIAN *et al*., 2014).

Além dos motivos apresentados, abrir dados é uma obrigação legal para a administração pública brasileira, conforme estipula a Lei nº 12.527/2011 (BRASIL, 2011). "A abertura de dados governamentais não se apresenta como mera alternativa de viabilização da transparência pública, mas como um dever a ser cumprido pelo administrador público" (TCU, 2015, n.p.).

A obrigatoriedade por lei é apresentada pelo TCU (2015, n.p.), que cita as seguintes leis, instrução normativa e decretos:

- Lei complementar 101/2000 (Lei de Responsabilidade Fiscal – LRF);
- Lei Complementar 131/2009 (Lei da Transparência);
- Decreto s/n de 15 de setembro de 2011, que instituiu o plano de ação nacional por meio do qual o Brasil, como um dos países que celebraram a Parceria para Governo Aberto (OGP);
- Lei 12.527/2011 (Lei de Acesso à Informação);
- Instrução Normativa SLTI/MP – 4/2012, que instituiu a Infraestrutura Nacional de Dados Abertos (Inda);
- Decreto 8.243/2014, que instituiu a Política Nacional de Participação Social – PNPS, com o objetivo de fortalecer e articular os mecanismos e as instâncias

> democráticas de diálogo e a atuação conjunta entre a Administração Pública Federal e a sociedade civil.

As possibilidades dos dados abertos governamentais vêm ao encontro do que o cidadão precisa para exercer a sua cidadania na sociedade em que vive, como a participação direta na tomada de decisões do Estado. Ao assumir papel ativo e seu poder de opinião, a mudança nas estruturas sociais são inevitáveis (MACIEL, 2008).

Com o avanço tecnológico e com os recursos das TICs inovando no meio governamental, surge a necessidade da democracia digital (ou e-democracia) e do governo eletrônico (do inglês e-gov ou *electronic government*). Cristiano Maciel (2008, p. 16) enfatiza a importância dos recursos tecnológicas como facilitadores para a participação social e explica:

> - O uso de TICs e de Comunicação Mediada por Computador (CMC) para intensificar a participação ativa dos cidadãos e dar suporte à colaboração entre os diversos atores, tais como cidadãos, governos, sociedade civil, entre outros, na elaboração de políticas públicas é chamada de democracia eletrônica.
> - Governo Eletrônico significa fundamentalmente as estratégias utilizadas pelo governo para uso dos recursos das Tecnologias de Informação e da Comunicação (TIC 's), com o intuito de modernizar a máquina administrativa e atender as necessidades do cidadão.

Na era do e-gov, o Estado tem a responsabilidade de modernizar-se, estreitando laços com o povo por meio da inovação tecnológica, promovendo a cultura da informação com as TICs, contribuindo para mudanças no cenário governamental e aumentando a capacidade de computação, estruturação e alcance da informação pública. A mudança de paradigma caracteriza-se pela descentralização do conhecimento e pela disseminação dos dados para a tomada de decisão e geração de novos conhecimentos.

Com a tecnologia provendo recursos e considerando a importância da informação no mundo moderno, é possível aprofundar o conceito de dados abertos seguindo por duas vias: entrando na área da ciência da informação na *web* e dando importância à qualidade, formatação e abrangência dos dados disponibilizados na rede mundial de computadores. É possível agregar valor (ou mais informação/conteúdo) aos dados publicados na internet (ou *world wide web*) criando conexões. Com esse enriquecimento dos dados surge a classificação baseada no conceito de dados conectados (do inglês *linked data*) de Tim Berners-Lee (2006, n.p.): "A *web* semântica não é apenas sobre colocar dados na web. É sobre fazer *links*, para que uma pessoa ou máquina possa explorar a *web* de dados. Com dados conectados, quando você tem alguns deles, pode encontrar outros dados relacionados."

Para possibilitar o reuso da informação e aumentar o alcance dos dados publicados na internet, é importante ponderar sobre o formato da publicação, visando a uma estrutura de arquivo com o potencial de interligação conforme o contexto ou área relacionada. Dados abertos conectados "[...] tratam de um conjunto de boas práticas para publicar e conectar

conjuntos de dados estruturados na *web*, formando assim uma *web* de dados" (PENTEADO; BITTENCOURT; ISOTANI, 2019b, p. 2).

O uso da *web* semântica nos dados gerados permite que agentes de *software* também consigam processar, compartilhar e "entender" as informações descritas pelos dados. De acordo com Isotani e Bittencourt (2015, p. 27),

> A *web* semântica estende a *web* clássica, provendo uma estrutura semântica para páginas *web*, a qual permite que tanto agentes humanos quanto agentes de software possam entender o conteúdo presente em páginas *web*. Dessa forma, a *web* semântica provê um ambiente em que agentes de software podem navegar através de páginas web e executar tarefas sofisticadas.

Outro conceito que surge nesse contexto da *web* semântica são as ontologias. Na ciência da computação, uma ontologia pode ser definida como (MIZOGUCHI, 2004 *apud* ISOTANI; BITTENCOURT, 2015, p. 95):

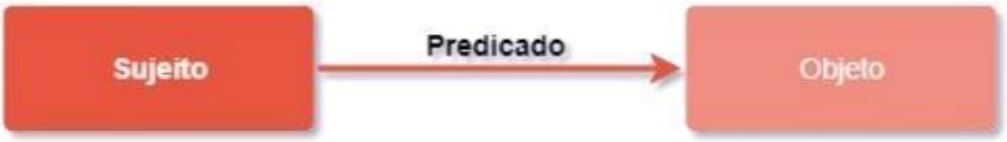
> Um conjunto de conceitos fundamentais e suas relações, que capta como as pessoas entendem (ou interpretam) o domínio em questão e permite a representação de tal entendimento de maneira formal, compreensível por humanos e computadores.

A concepção e o uso das ontologias para representar o conhecimento e descrever os dados fazem parte da ideia da *web* semântica e são recomendadas por Tim Berners-Lee (ISOTANI; BITTENCOURT, 2015). Ontologias de domínio e de tarefa são necessárias para criar sistemas mais flexíveis e inteligentes que possam ser aplicados em diversos domínios (ISOTANI; BITTENCOURT, 2015). Para isso, existem linguagens de representação formal das ontologias processáveis por computador (por exemplo RDF). Isotani e Bittencourt (2015, p. 104) definem *Resource Description Framework* - RDF como:

> RDF é a especificação proposta pelo *World Wide Web Consortium* (W3C) para descrever metadados. Ela permite criar triplas que contêm um nó sujeito, uma relação chamada de predicado e o nó objeto (sujeito, predicado, objeto). Mediante essa tripla, é possível indicar a relação entre dados e usá-la para representar a semântica contida neles.

Na figura 1 temos a representação gráfica de uma tripla RDF com o sujeito (ou recurso), predicado (propriedade) e objeto (ou valor):

FIGURA 1 – GRAFO RDF COM TRIPLA SUJEITO, PREDICADO E OBJETO



Fonte: Silva (2018)

Com base nos conceitos apresentados, os quatro princípios de dados conectados, introduzidos por Tim Berners-Lee (2006, n.p.) são:

> I. Usar URIs como nome para recursos;
>
> II. Usar URIs HTTP para que as pessoas possam encontrar esses nomes;
>
> III. Quando uma URI for acessada, garantir que informações úteis possam ser obtidas por meio dessa URI, as quais devem estar representadas no formato RDF;
>
> IV. Incluir links para outras URIs de forma que outros recursos possam ser descobertos

Onde RDF é um modelo padrão para representação de dados na *web* (ligando coisas arbitrárias descritas por RDF) e URIs (*Uniform Resource Identifier*) são mecanismos de identificação global e único. Nos termos de Lóscio *et al*. (2018, p. 19),

> A *web* de documentos é baseada em um conjunto de padrões, incluindo: um mecanismo de identificação global e único, os URIs (*Uniform Resource Identifier*); um mecanismo de acesso universal, o HTTP; e um formato padrão para representação de conteúdo, o HTML. De modo semelhante, a *web* de dados tem por base alguns padrões, como: o mesmo mecanismo de identificação e acesso universal usado na *web* de documentos (URIs e HTTP, respectivamente) e um modelo padrão para representação de dados, o RDF.

Existem diversos formatos de arquivos para a representação dos dados baseados nos princípios do padrão RDF. É possível serializar os dados utilizando formatos como JSON-LD, RDFa (código RDF embutido em HTML), RDF/XML etc. Considerando a diversidade de arquivos para a estruturação de dados, é possível classificar os dados abertos de acordo com uma escala baseada em 5 estrelas, proposta por Tim Berners-Lee (BERNERS-LEE, 2006, n.p.). No sistema de classificação por estrela, designado "Sistema de 5 Estrelas", atribui-se uma estrela ao dado que atenda a estas duas características: publicado na web (em qualquer formato: arquivo, imagem, tabela ou documento) e associado a uma licença que permita o seu

uso e reuso sem restrições. A escala completa é apresentada a seguir (LÓSCIO *et al*., 2018, p. 14):

> 1. ★ Disponível na web com licença aberta;
> 2. ★★ O dado deve estar em formato estruturado, legível por máquina;
> 3. ★★★ Formato estruturado e aberto (exemplo: CSV que não depende de programas proprietários para ser manipulado);
> 4. ★★★★ Todos os itens acima e usam padrões abertos para identificar dados (exemplo: identificadores URI), permitindo criar ligações com os dados;
> 5. ★★★★★ Dados conectados com outros dados (cria-se um contexto ao dado, vinculando-se com outros dados).

A carência de critérios e padrões globais para representação de dados abertos equivalentes, aliada a enorme disponibilidade de arquivos (pdf, csv, xml, html, json, formatos proprietários etc.) para publicações na internet, constitui um problema desafiador para o potencial transformador dos dados abertos, principalmente na esfera governamental. Pode-se afirmar que dados abertos governamentais vêm sendo publicados há alguns anos, porém os órgãos governamentais possuem autonomia na escolha sobre como e em que formato os dados serão publicados. A falta de padronização ou formatos processáveis por computador gera dificuldades no reuso das informações por outras pessoas ou aplicações, como citado por Penteado, Maldonado e Isotani (2021), Penteado, Bittencourt e Isotani (2019b), Alcantara *et al*. (2015) e Bandeira *et al*.(2015).

Trabalham contra a propagação da informação para uso futuro na *web* os dados publicados em diversos formatos. Os diferentes formatos das publicações e a ausência de dados legíveis por máquinas não só impedem a criação de novos serviços e produtos baseados em dados como dificultam o alcance do objetivo dos dados abertos: serem acessíveis para qualquer cidadão ou agentes de *software* (PENTEADO; BITTENCOURT; ISOTANI, 2019a). Alcantara *et al*. (2015, n.p.) assim se posicionam sobre a diversidade dos dados publicados e seu consumo:

> O consumo de dados abertos pode ser feito por públicos de diferentes perfis. Desenvolvedores tendem a utilizar APIs e *Webservices*, pois facilitam o consumo automatizado sob os dados. Por outro lado, pessoas que não são desenvolvedores preferem que a base esteja em formatos mais comuns, como PDF, CSV e DOC, pois são formatos legíveis para eles

O uso de dados abertos conectados (DAC) é uma alternativa encontrada pela comunidade *web* para o problema da baixa qualidade dos dados e o da problemática da distribuição (ALCANTARA *et al*., 2015). Do ponto de vista tecnológico da abordagem dos

dados abertos conectados (do inglês *linked open data* - LOD), Penteado, Bittencourt e Isotani (2019b, p. 2) destacam as seguintes vantagens:

> Os dados abertos conectados (LOD) são um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na *web*, formando assim uma "*web de dados*", reaproveitando a infraestrutura da *web* e também tecnologias semânticas já consolidadas.

Com o uso de DAC e ontologias as limitações derivadas da falta de padronização minimizam e as conexões permitem o reuso e a geração de novos conhecimentos derivados dos dados disponibilizados. Vejamos o cenário nacional relatado por Penteado, Maldonado e Isotani (2021):

> A descentralização das ações de liberação de dados faz com que os conjuntos de dados sejam produzidos em diferentes formatos, gerando silos de informações, que isolam os dados de outras fontes e impede que se crie conexões para atender a consultas de dados mais complexas. Essa limitação restringe o potencial de reuso das informações, pois dificulta o cruzamento de informações de diferentes fontes, já que o processo de conhecimento das bases, limpeza dos dados, descoberta de conexões e interligação com outras bases de dados fica a cargo de quem os consome.

A utilização dos dados abertos conectados para interligar as informações publicadas, criando um contexto e contribuindo para a geração de conhecimento com qualidade e com maior potencial para a tomada de decisão e consumo surge como uma possibilidade para o problema da variedade de formatos disponíveis oriundos de diversas fontes ou órgãos públicos (PENTEADO; MALDONADO; ISOTANI, 2021).

2.1 Aplicabilidade dos conceitos no cenário nacional

Para minimizar os problemas decorrentes da falta de padronização e da diversidade de formatos das publicações, foram criadas diretrizes para a produção e formatação de dados abertos no âmbito nacional, a exemplo dos Padrões de Interoperabilidade de Governo Eletrônico ou e-Ping (EPING, 2021) e das práticas estabelecidas pela Infraestrutura Nacional de Dados Abertos (INDA) (DADOS.GOV.BR, 2021). Tais padrões são especificações técnicas que possibilitam não só as condições de interação entre sistemas como regulamentam a utilização das TICs no governo federal, "[...] sendo o uso de vocabulários e ontologias recomendado no e-PING (arquitetura que define as políticas e especificações técnicas que regulamentam as TIC nos serviços de governo eletrônico no âmbito federal, como forma de incentivar a interoperabilidade entre diferentes entidades da Federação)" (PENTEADO; BITTENCOURT; ISOTANI, 2019a). Quanto à produção dos dados, existem as práticas de especificação de metadados da INDA, que considera os formatos de arquivos, aderência ao e-PING, metadados, disponibilidade de URL, entre outros (PENTEADO, 2020).

Propostas para produção, consumo, formatação e representações de dados abertos públicos em diferentes contextos são abundantes na literatura no âmbito nacional. As pesquisas empreendidas retornaram modelos para dados públicos legislativos (BRANDT; VIDOTTI; SEGUNDO, 2018), censitários (censo da educação básica, da educação superior, censo demográfico etc.) (CARVANO, 2018) e educacionais (plataforma Lattes, dados das instituições sobre professores, projetos, cursos, indicadores educacionais) (MORAES NETO *et al.*, 2020) (PENTEADO; BITTENCOURT; ISOTANI, 2019a) (ALENCAR *et al.*, 2018) (RAUTENBERG *et al.*, 2017). Existem entidades públicas que são provedores de informações, estatísticas e indicadores nacionais como o INEP, IBGE, plataforma Sucupira, CAPES, Portal Brasileiro de Dados Abertos, universidades, entre outros. Essas entidades produzem regularmente informações sobre o sistema educacional, informações geográficas e estatísticas, os programas de pós-graduação etc. Os dados produzidos por estes órgãos governamentais são fontes de estudos na área dos dados abertos com diversas propostas de modelos e métodos para extração, representação ou mapeamento dos dados brutos para dados abertos governamentais conectados.

A título de exemplo, tomemos o modelo de dados abertos conectados para um conjunto de dados legislativos da câmara dos deputados, apresentado por Brandt, Vidotti e Segundo (2018). Nesta pesquisa foi selecionado o conjunto de dados "deputados" (formado por partido político, unidade federativa, e-mail, legislatura etc.) para a estruturação dos dados em RDF e reuso de vocabulários e padrões já estabelecidos na web semântica (Dublin Core, Friend of a Friend - FOAF, RDF e RDF Schema), inclusive os vocabulários de áreas correlatas (ontologia da Câmara dos Deputados italiana e a da Assembleia Nacional Francesa). A fonte dos dados foi o portal Dados Abertos da Câmara dos Deputados (BRANDT; VIDOTTI; SEGUNDO, 2018).

Inserida no quesito censo, está a dissertação de mestrado de Carvano (2018), que consome os dados abertos do INEP e do IBGE para a construção de um repositório de estatísticas educacionais públicas brasileiras com a finalidade de agregar valor às informações para a tomada de decisão. Além da proposta do modelo para representação dos dados, o objetivo principal do projeto é construir uma solução para mapeamento dos dados brutos em indicadores sociais.

Passando para o contexto educacional, foco deste presente trabalho, vejamos como os dados abertos deste cenário são descritos por Penteado, Bittencourt e Isotani (2019a, p. 175):

> Os dados abertos educacionais trazem informações importantes sobre o cenário educacional de um país. Sua publicação traz impactos tanto em transparência quanto no aumento do potencial econômico para a sociedade como um todo por meio da gestão da aprendizagem e da tomada de decisão baseada em evidências.

Bandeira *et al*. (2015, p. 48) também ressaltam a importância dos dados abertos da educação ao afirmarem que:

> No contexto educacional, os dados abertos são importantes em diversas atividades, como por exemplo, podem ser utilizados no desenvolvimento de soluções tecnológicas que auxiliem na tomada de decisão de professores e gestores escolares, bem como na ampliação da oferta e produção de novos conhecimentos que sirvam de base para o desenvolvimento e aprimoramento de recursos educacionais.

Penteado, Bittencourt e Isotani (2019a, p. 178) falam dos benefícios da publicação dos dados educacionais nos seguintes termos:

> Os dados educacionais podem ser usados para diferentes finalidades, como: planejamento de metas e objetivos a serem alcançados pelos gestores de uma região; avaliar a efetividade de medidas adotadas no contexto educacional; desenvolvimento de pesquisas; produtos de empresas que atuam no mercado educacional, buscando trazer tanto benefício econômico quanto de transparência, e servindo de base para a avaliação e proposição de melhorias no sistema educacional.

Da mesma forma que os DAGs, os dados educacionais também são publicados em uma grande variedade de formatos. Além disso, os órgãos possuem autonomia de escolha sobre como formatar e representar semanticamente os dados abertos. Para trazer os benefícios a que a sociedade tem direito (transparência, difusão dos dados públicos para tomada de decisão ou para acompanhamento do cenário educacional), os DAEs devem ser produzidos e publicados de acordo com os padrões conhecidos da *web*. A falta de padronização e a diversidade de formatos também são encontradas no contexto educacional como bem observam Bandeira *et al*. (2015, p. 48):

> Observa-se que os dados educacionais são publicados, predominantemente, em formato não estruturado – o que limita a sua descrição e reutilização. Assim, o processo de consumo tem ocorrido mediante muito esforço e custo devido à baixa qualidade dos dados disponibilizados.

"Na educação, existe uma diversidade de dados sendo gerados todos os dias. Contudo, essas bases apresentam problemas que dificultam o enriquecimento e a conexão dos dados neste cenário" (ALCANTARA *et al*., 2015, n.p.). Para corrigi-los, esta seção terá como escopo o levantamento de soluções que se proponham a suprir a falta de representações de qualidade dos dados abertos educacionais no cenário nacional bem com o uso da *web* semântica visando à minimização de tais problemas. A pesquisa foi dividida em duas frentes: modelos para representação dos dados e ferramentas ou aplicações tecnológicas para geração de dados abertos no Brasil.

Para a educação superior, uma variedade de estudos foi encontrada. Conforme Rautenberg *et al*. (2017, p. 118), "as universidades são organizações consumidoras, produtoras e disseminadoras de conhecimento[...]Organizar, formalizar e compartilhar indicadores sobre o conhecimento produzido e disseminado é uma tarefa desafiadora a estas instituições".

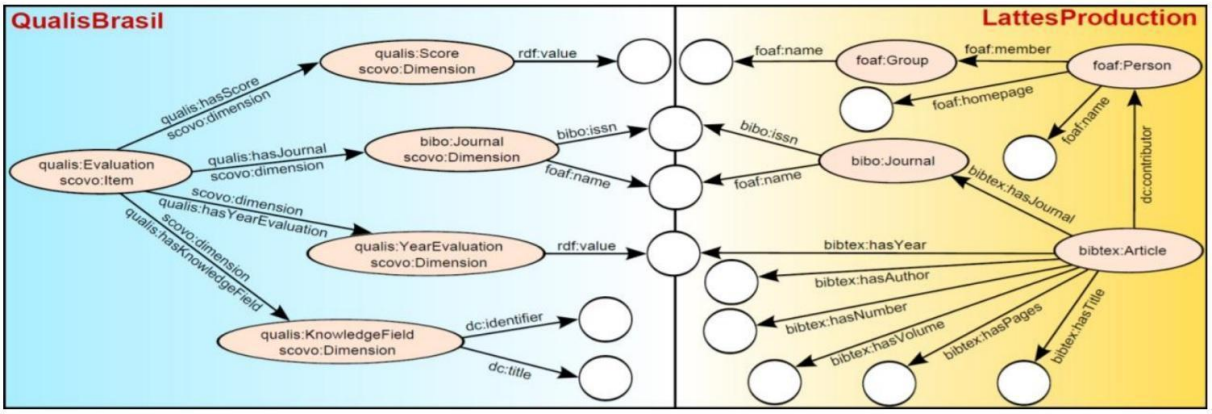
Promovidas por órgãos públicos, duas importantes bases de dados abertos desempenham papel fundamental como fonte de dados da educação superior: a plataforma

Lattes e a Sucupira. O trabalho de Rautenberg *et al*. (2017) consome os dados dessas fontes em estudos de caso cientométricos em uma universidade brasileira. Os dados são tratados, organizados e cruzados a fim de produzir conhecimento contextualizado em uma universidade pública. Para elevar os dados abertos do índice Qualis e da plataforma Lattes à quinta estrela (dados conectados), alguns vocabulários e ontologias disponíveis na *web* foram utilizados na representação dos dados abertos (RAUTENBERG *et al*., 2017, p 131), a saber:

- SCOVO (The Statistical COre VOcabulary): é um vocabulário simples para representar dados estatísticos na web. Neste trabalho, é usado para organizar o índice Qualis na forma multidimensional;
- DC (Dublin Core): é um vocabulário amplamente utilizado para descrever recursos. É utilizado para: a) melhor representar as áreas de conhecimento no grafo QualisBrasil (elementos dc:identifier e dc:title); e b) relacionar um indivíduo a um artigo científico como um coautor (dc:contributor);
- BIBO (Bibliographic Ontology Specification): é uma ontologia que modela os conceitos e as propriedades de referências bibliográficas. Seus elementos são usados para representar os periódicos (journals) nos grafos;
- FOAF (Friend-of-a-Friend): é um vocabulário utilizado para relacionar entidades a informações na web. No grafo Lattes Production, por exemplo, mapeia grupos (cursos, departamentos ou centros) a seus membros;
- BIBTEX9 (Transformation of bibTeX into an OWL ontology): é uma ontologia que define os elementos de referências bibliográficas. É usada para mapear as referências capturadas da plataforma Lattes.

A figura 2 ilustra como os dados são representados no modelo RDF.

FIGURA 2 – REPRESENTAÇÃO DOS GRAFOS QUALISBRASIL E LATTESPRODUCTION



Fonte: Rautenberg *et al*. (2017)

A Unidade Acadêmica de Informática (UAI) do IFPB e a Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) são algumas das entidades públicas que

optaram pela criação das próprias ontologias e soluções, respectivamente a ontologia OUAI (ALENCAR et al., 2018) e o Chatbot. Este último tem a finalidade de disponibilizar o catálogo de cursos da instituição (MORAES NETO et al., 2020).

Alencar *et al*. (2018) apresentam uma ideia para publicar dados em formato aberto de forma que estes possam ser consumidos tanto por humanos quanto por software no escopo da UAI do IFPB. O conjunto de dados engloba informações institucionais de interesse público(cursos ofertados e as disciplinas associadas, informações sobre professores, projetos, cursos e áreas de atuação da UAI). Para referenciar semanticamente os dados, foi desenvolvida a ontologia de domínio OUAI (*Ontology for Universities and Academic Information*), que reusa termos de vocabulários recomendados e acrescenta outros específicos. Para o consumo dos dados, foi especificada a aplicação *web* OpenUAI (ALENCAR *et al*., 2018). As informações dos cursos e matrizes curriculares foram extraídas do sistema de controle acadêmico para planilhas CSV. A conversão dos dados para RDF é feita com a utilização dos termos da ontologia OUAI (ALENCAR *et al*., 2018). No quadro 1, é apresentado um exemplo de mapeamento entre campos e termos da OUAI.

QUADRO 1 – MAPEAMENTO ENTRE CAMPOS E TERMOS DA OUAI

Campo do CSV	Termo da OUAI	Tipo do Termo	Tipo de Dado
Disciplina	dc:title	Propriedade de dados	String
Período	ouai:courseSemester	Propriedade de dados	Inteiro
Carga Horária	time:hours	Propriedade de dados	Inteiro
Ementa	ouai:courseContent	Propriedade de dados	String
Bibliografia básica	ouai:hasBibliography	Propriedade de objetos	Recurso: ouai:bibliography

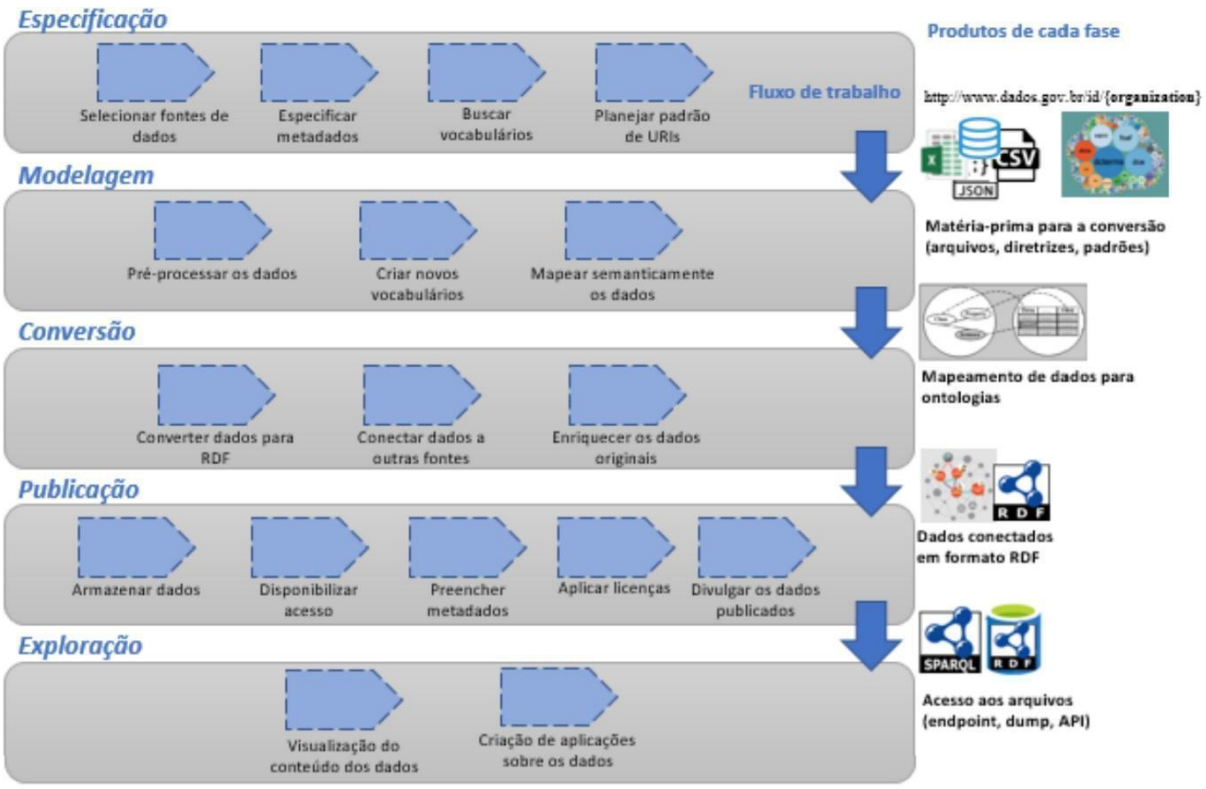
Fonte: Alencar *et al*. (2018)

Deseja-se com a fundamentação deste capítulo buscar referências de modelos de dados de acordo com as especificações do governo federal. Para o alcance do objetivo, pretende-se adotar padrões para representar os dados educacionais mediante o uso de especificações e ontologias cujas anotações semânticas possibilitem a troca de informações e as integrações entre diferentes sistemas. Dos estudos de Penteado (2020) e de Penteado *et al*. (2019c; 2019b) vêm a referência e inspiração para a realização deste trabalho, respectivamente quanto a metodologia para a geração dos dados com uso de ontologia e vocabulários para o contexto educacional, o modelo de referência para dados abertos educacionais em nível macro e o

metaprocesso para a transformação de dados educacionais em dados conectados. Tais propostas inspiraram a construção do método e o modelo de dados aqui empregado, evitando a necessidade da definição de uma ontologia própria para a UFMT. Entre as três propostas, a de Penteado (2020) foi a principal referência para este trabalho porque apresenta um modelo completo de infraestrutura para publicação de dados abertos governamentais conectados de qualidade e por sua produção científica sobre dados abertos no Brasil, além do fato de ele ser um pesquisador com inúmeras publicações e citações na área de dados abertos educacionais, ambiente desta dissertação.

O metaprocesso proposto por Penteado, Bittencourt e Isotani (2019b) é composto por cinco fases: especificação, modelagem, conversão, publicação e exploração. No trabalho mais recente sobre o tema (PENTEADO; MALDONADO; ISOTANI, 2021), os autores adicionam uma sexta fase: manutenção. Para efeito de escopo deste trabalho, serão usadas as quatro fases iniciais: extração dos dados das diversas bases de dados, respectivo processamento e conversão para os modelos adotados e a publicação dos dados no catálogo da instituição. Na figura 3 está representado o esquema do metaprocesso.

FIGURA 3 – METAPROCESSO PARA PUBLICAÇÃO DE DADOS ABERTOS CONECTADOS



Fonte: Penteado, Bittencourte e Isotani (2019b)

O metaprocesso apresenta instruções ou um guia com os passos importantes do processo de transformação dos dados educacionais. A fase de especificação serve para "levantar quais os dados serão publicados em formato conectado, os metadados obrigatórios e opcionais que ele deve conter, os vocabulários (ou ontologias) que deverão descrever esses dados e o padrão de URI a ser adotado pela organização" (PENTEADO; BITTENCOURT; ISOTANI, 2019b, p. 4). É nesta fase que será definida a fonte de origem dos dados e a seleção dos vocabulários e metadados para representação e enriquecimento semântico dos dados educacionais. Considerando que uma IES pode possuir várias bases de dados e sistemas, deve-se atentar para duas partes importantes do processo: a seleção criteriosa dos dados públicos que se quer coletar, a fim de garantir a proteção dos dados confidenciais, e a escolha das ontologias.

Também pertence à fase de especificação a definição dos metadados obrigatórios e opcionais. No portal do governo federal (DADOS.GOV.BR, 2021, n.p.) encontra-se a seguinte definição: "metadados são dados sobre os dados, ou seja, são informações que possibilitam organizar, classificar, relacionar e inferir novos dados sobre o conjunto de dados". A seguir, o conjunto de metadados obrigatórios e desejáveis extraídos do portal.

Os obrigatórios são:

- **Título**: Nome do conjunto de dados.
- **Descrição**: Uma breve explicação sobre os dados.
- **Catálogo origem**: Página (URL) do órgão onde está publicado o conjunto de dados.
- **Órgão responsável**: Nome e sigla do órgão ou entidade responsável pela publicação do conjunto de dados.
- **Categorias no VCGE**: O Vocabulário Controlado de Governo Eletrônico é uma lista hierarquizada de assuntos do governo que utiliza termos comuns e é voltada para a sociedade. Para navegar e escolher as categorias acesse o VCGE em https://vocab.e.gov.br/2011/03/vcge.
- **Recursos**: Um conjunto de dados pode ser composto por mais de um arquivo de dados. O critério básico para separar vários recursos em mais de um conjunto de dados é a constatação de que eles divergem em vários metadados.

  Identificador: URL persistente que aponta para o recurso na Web.

  Título: Nome do recurso.

  Formato: Formato do recurso. Ex.: XML, JSON, CSV, etc.

  Descrição: Breve detalhamento sobre o conteúdo do recurso.

E os desejáveis são:

- **Etiquetas**: Lista de palavras chaves relacionadas ao conjunto de dados, e que são úteis na classificação e busca dele.
- **Autoria**: Instituição ou pessoa responsável pela produção do recurso.

- **Documentação**: URL de documento que expõe detalhes sobre o conjunto de dados.
- **Cobertura geográfica**: Localização ou região geográfica a que se referem os dados. Ex.: Recife.
- **Cobertura temporal**: Data ou período à que referem os dados. Ex.: 03/2012.
- **Granularidade geográfica**: Precisão geográfica da cobertura geográfica. Ex.: municipal.
- **Granularidade temporal**: Precisão temporal da cobertura temporal. Ex.: mês.
- **Frequência de atualização**: Frequência temporal com que o conjunto de dados é atualizado.
- **Referências**: Relações com outros conjuntos de dados.
- **Metodologia**: Processo de criação dos dados.
- **Vocabulário/ontologia**: Documentos estruturados com metadados específicos do conjunto de dados.

Avançando na descrição das fases do metaprocesso, chegamos à modelagem, momento em que os dados do passo anterior são processados, o mapeamento dos campos oriundos das fontes de origem para gerar os dados das ontologias é feito e os termos e vocabulários são escolhidos para representação. Penteado, Bittencourt e Isotani (2019b, p. 5) assim descrevem os passos da modelagem:

> Pré processar os dados, de modo a filtrar dados de baixa qualidade (faltando, mal formatado, etc.). Em seguida, são levantadas as necessidades de anotações semânticas que ainda não foram atendidas pelos vocabulários existentes e são criados novos vocabulários. Assim, o próximo passo é anotar os dados originais com sua respectiva anotação semântica. Com todos os mapeamentos feitos (produto desta fase), se inicia a fase de conversão.

O próximo passo do metaprocesso é a conversão. Etapa em que os dados originais, já mapeados, são transformados para o modelo de dados semântico (formato RDF) e a saída é apresentada em um grafo RDF (PENTEADO; BITTENCOURT; ISOTANI, 2019b). Além da conversão dos dados para RDF, também ocorrem a conexão com outros dados e a apresentação do produto final: dados conectados em formato RDF.

Por fim, temos a etapa da publicação dos dados mapeados e convertidos que serão disponibilizados para consulta pública. A divulgação dos dados abertos gerados pode ocorrer das seguintes maneiras: publicação no portal da instituição, *web services* ou em catálogo/repositório de dados abertos como o CKAN ou o Socrata. Penteado, Bittencourt e Isotani (2019b, p. 5) deixam claro que

> Nela o conteúdo do grafo gerado é armazenado para uso, seja por meio de *upload* em um *website* de catálogo de dados (ex.: CKAN), um arquivo *dump* ou de um *endpoint* SPARQL. Além disso, deve-se divulgar o ponto de acesso a essa fonte de dados, seja a URL do arquivo ou do *endpoint* SPARQL. Por fim, esse grafo deve ser versionado, preenchidos seus metadados e a ele deve ser atribuída uma licença de uso.

Com a disponibilização dos DAEs em uma ferramenta para catalogação de dados abertos é possível utilizar os recursos do catálogo, permitindo a consulta por agentes de *software* ou pessoas e para organização dos dados públicos (como uma url única para identificação e acesso). CKAN é um *software* livre para catálogos de dados que disponibiliza para os usuários finais o acesso e a consulta aos dados abertos publicados. É uma plataforma *web*, projetada pela Open Knowledge Foundation (OKF) para a publicação e compartilhamento de dados abertos (REIS *et al*., 2019), que possui *interface* para compreensão por humanos e os recursos de acesso via API do próprio CKAN (o que possibilita a comunicação com outros sistemas). Com o resultado publicado enriquecido semanticamente também é possível realizar a pesquisa no catálogo de dados abertos conectados via *interface* da API, assim como viabilizar o uso do conjunto de dados pela sociedade em geral (MARTINS, 2018).

O uso de APIs (*Application Programming Interface* ou Interface de Programação de Aplicações) possibilita a integração de serviços, a publicação e o gerenciamento dos dados no catálogo das instituições mediante o uso de uma interface *web*. A API é um recurso poderoso para fazer com que diferentes soluções emerjam da comunicação entre sistemas distintos. Ela pode ser usada como fonte de dados, como demonstrado por Torino, Trevisan e Vidotti (2019). O material de análise dos autores é composto por um conjunto de dados acerca das avaliações das pós-graduações stricto sensu, disponíveis no portal de dados abertos da CAPES e consumidos via API.

Exemplos de aplicações dos recursos da API mencionadas no estudo: o portal de dados abertos da CAPES utiliza o software CKAN, que disponibiliza uma API baseada em JSON; os conjuntos de dados analisados estão organizados por nomes associados ao hiperlink individual de cada conjunto; os recursos de API do CKAN também se destinam à publicação dos dados no catálogo, do mesmo modo como é feito pela aplicação UnBGOLD que, após o enriquecimento semântico, publica via API o conjunto de dados no catálogo CKAN da UnB (MARTINS *et al*., 2018).

2.2 Casos empíricos

Esta seção apresentará trabalhos que, por afinidade de propósito, subsidiaram a concretização dos objetivos da presente pesquisa: extrair as informações dos bancos de dados e sistemas e automatizar os processos de conversão e publicação no catálogo de dados abertos de suas instituições, mediante enriquecimento semântico e uso de TICs.

Na literatura, há pesquisadores que desenvolveram suas próprias soluções tecnológicas. Tomando a Universidade Federal do Maranhão (UFMA) por início da exposição, em cuja resolução foram encontrados diferentes métodos, eis a solução adotada por Silva *et al*. (2016) para a extração dos dados dos diversos portais da universidade: uso das

técnicas de raspagem de dados (*data scraping*) e *deep web*, com a criação de algoritmos próprios para a coleta das informações e modelagem dos dados seguindo o padrão utilizado pela *Open University*.

Oliveira, Guimarães e Costa (2018) também apresentaram uma proposta de migração de dados abertos para dados conectados para a Universidade Federal do Maranhão. A solução é baseada em quatro etapas: modelagem (das entidades que são mapeadas como recursos RDFs), extração e geração de dados (através de alguns códigos escritos na linguagem Python para extração dos dados para o formato CSV que serviram de entrada para o *software* Open Refine), armazenamento (e um servidor *triple store* Apache Jena Fuseki) e acesso (com consultas na linguagem SPARQL).

A linguagem Python apareceu em variados estudos, sendo empregada para diferentes finalidades. Na fase de extração dos dados, merecem destaque dois estudos: o de Alencar *et al*. (2018, p. 139), pela descrição de como "[...] foram coletados os currículos Lattes dos professores lotados na UAI e os dados extraídos eram originalmente arquivos XML obtidos através de um *scraper* codificado na linguagem Python" e o de Silveira (2021, p. 59), pela justificativa dada "[...] para atender à atividade de extração de relações, foi codificado um script em Python que realiza a leitura de um arquivo com um conjunto de sentenças". Na fase de transformação dos dados, temos a proposta de Moraes Neto *et al*. (2020), da qual extraímos as seguintes informações: a aplicação Python é responsável pelo Processamento de Linguagem Natural (PLN), treinamento do modelo de aprendizagem e classificação dos textos de entrada. Quanto à carga dos dados, foi assim mencionada por Carvano (2018): scripts na linguagem Python foram utilizados para auxiliar o processo de carga dessas informações no banco de metadados e para o processamento dos dados para a produção de indicadores sociais.

No entanto, uma outra forma de solucionar o problema da transformação dos dados foi encontrada durante a pesquisa: a abordagem ETL (*Extract-Transform-Load*, que em português significa extração, transformação e carga dos dados). Alguns autores, ao invés de implementar scripts próprios utilizando-se de linguagens de programação, optaram por criar soluções baseadas no conceito de ETL: integrar os dados da fonte de origem e submetê-los à transformação para que depois sejam carregados na fonte de destino (PENTAHO, 2021).

Entre as ferramentas de ETL disponíveis, é possível citar as seguintes: Oracle Data Integrator (ODI), Microsoft Integration Services (SSIS), os *open source* Talend e o Pentaho Data Integration (também conhecido por Kettle). Com elas é possível extrair os dados de origem de diversas fontes, realizar os mapeamentos para os termos e ontologias e formatar os dados para publicação utilizando os componentes parametrizáveis das ferramentas. O processo de transformação dos dados abertos governamentais pode ser conduzido por um fluxo (ou *workflow*) ETL. Os trabalhos apresentados a seguir utilizaram a abordagem ETL para orquestrar o processo de extração dos dados abertos e públicos, o seu mapeamento para

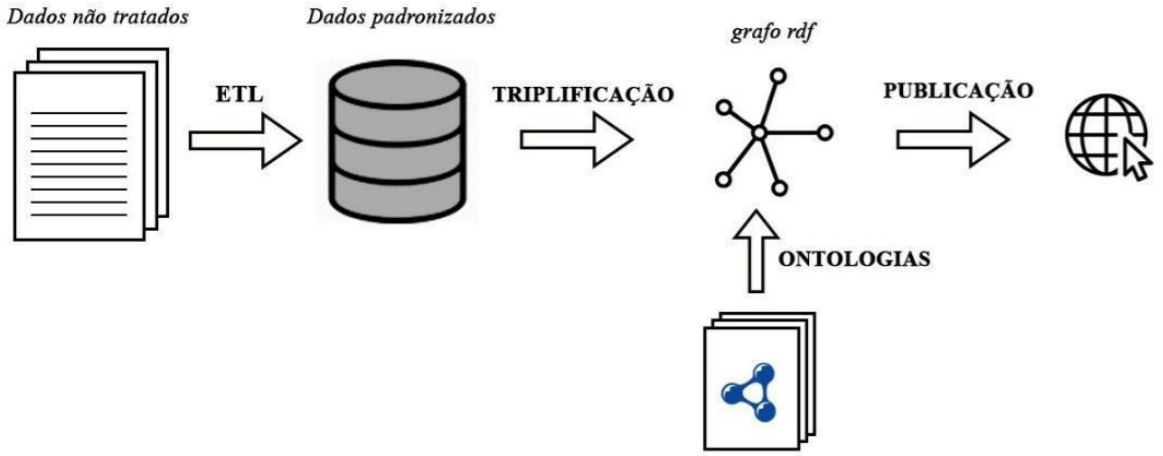
dados conectados (triplificação dos dados) e a sua publicação em um servidor ou catálogo de dados abertos institucionais.

No caso da Universidade de Brasília (UnB), uma proposta de arquitetura de publicação automatizada de dados abertos conectados foi materializada por meio da implementação de uma ferramenta denominada UnB Government Linked Open Data (UnBGOLD) (REIS *et al*., 2019; MARTINS, 2018; MARTINS *et al*., 2018). Por meio de um processo de ETL, os dados são extraídos das bases de dados dos seguintes sistemas da universidade: Sistema de Pessoal (SIPES), Sistema de Graduação (SIGRA), Sistema de Pós-Graduação (SIPPOS) e Sistema de Extensão e Pós-graduação (SIEX) – análoga realidade se verifica na UFMT. O UnBGOLD realiza a indexação semântica dos dados utilizando um vocabulário controlado e publica automaticamente os dados no CKAN (REIS *et al*., 2019; MARTINS, 2018; MARTINS *et al*., 2018).

A abordagem ETL, conduzindo o *workflow* de geração e publicação dos dados abertos conectados, também foi encontrada em estudos da Universidade Federal do Rio de Janeiro (SILVA, 2018) e do Instituto Militar de Engenharia (SILVEIRA, 2021). Em ambos os casos, a ferramenta Kettle da PENTAHO foi utilizada para implementar o fluxo de publicação. São características positivas do Kettle: ser código aberto, possuir uma comunidade de usuários consolidada e documentação e principalmente por dispor de extensões (ou *plugins*) para trabalhar com dados abertos conectados (SILVEIRA, 2021; SILVA, 2018).

Resumidamente, o ciclo de vida dos dados abertos conectados é transformar os dados limpos (entrada) em dados conectados (saída). Inicialmente os dados não tratados, extraídos das bases de dados da instituição, serão submetidos a um pré-processamento e depois transformados em RDF num processo conhecido como triplificação. O processo de ETL é responsável por realizar uma série de operações que permitem o mapeamento e enriquecimento semântico dos dados brutos. O fluxo da abordagem ETL para geração de dados abertos pode ser visto na figura 4:

FIGURA 4 – ESQUEMA GERAL DO MÉTODO ETL PARA GERAÇÃO DE DADOS ABERTOS



Fonte: Silva (2018)

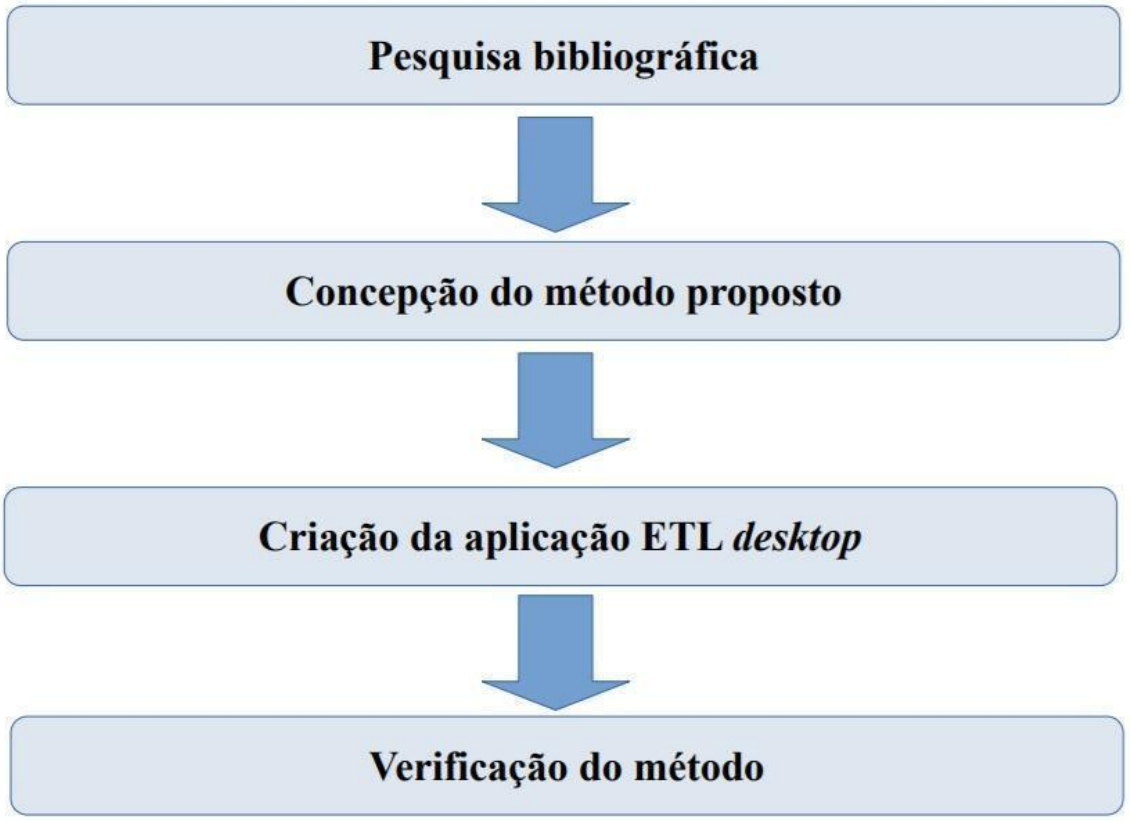
Com o foco fixado nas IES, nesta seção traçaram-se os caminhos que levaram à busca de informações detalhadas sobre DAG e à pesquisa sobre o modo como os DAGs são gerados e publicados pelas entidades do governo. Aliar o que foi apurado em relação ao que as universidades estão fazendo para abrir os dados de suas instituições ao conhecimento depreendido do aporte teórico da pesquisa, resultou em inspiração para criar o método proposto neste trabalho. Na literatura visitada, também foi possível encontrar modelos, vocabulários e ontologias para representar os dados abertos conforme o contexto e formas para agregar valor e aumentar a potencialidade de reuso dos DAG. Reiterando, o embasamento teórico, além de fundamental para a qualidade científica do trabalho, foi o estopim para proposição de um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT na abordagem ETL.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa aplicada foi a metodologia adotada no presente trabalho, cujo foco esteve concentrado na busca por inspiração para resolver o problema de gerar dados abertos educacionais de qualidade no contexto da UFMT. A pesquisa aplicada objetiva gerar conhecimento com finalidade de aplicação (PRODANOV; FREITAS, 2013 apud MACHADO et al., 2016).

O fluxo de trabalho (figura 5) consta das seguintes etapas:

FIGURA 5 – FLUXO DE TRABALHO



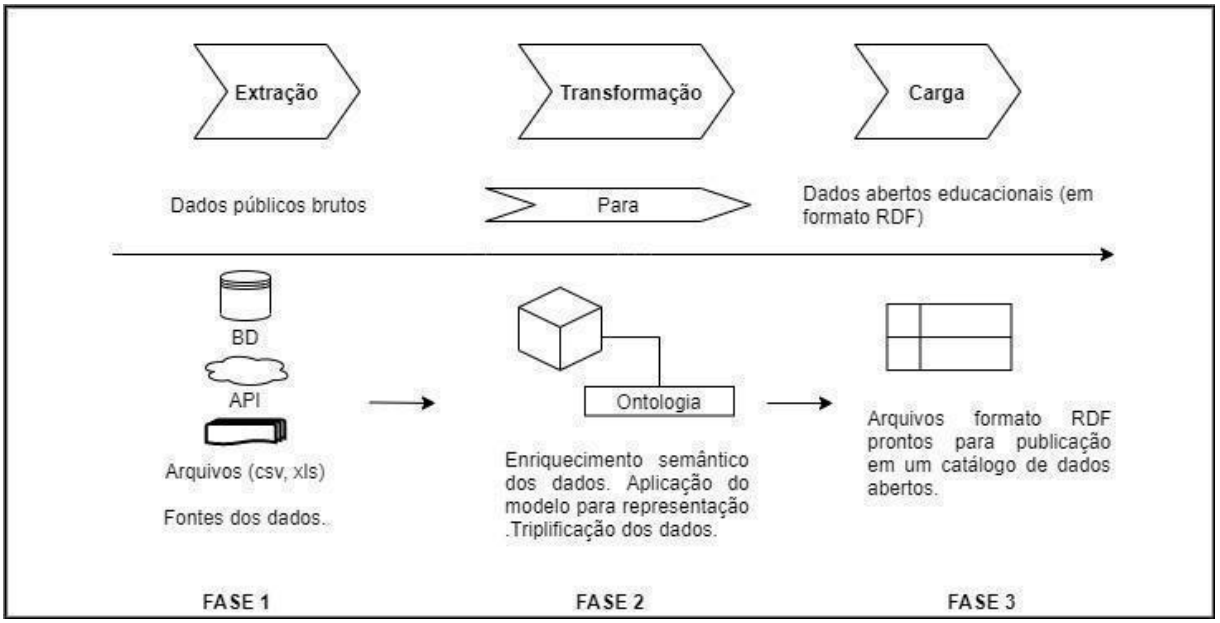
Fonte: Elaborado pelo autor (2021)

- **Pesquisa bibliográfica**: É a base para a fundamentação teórica, compreensão do problema e levantamento do estado da arte e dos trabalhos correlacionados. Caracteriza a estratégia desta pesquisa a busca por soluções e exemplos (ferramentas, modelos, processos, métodos ou *workflow*) para geração de dados abertos governamentais. O início da busca no portal de periódicos da CAPES dá-se pelo parâmetro assunto nos seguintes termos: ((dado* AND aberto* AND governamenta*) AND (m*todo* OR modelo* OR workflow OR ferramenta*)). Em seguida, após análise dos resultados obtidos, a pesquisa é aprofundada só

que agora com foco no cenário educacional, buscando referências de ontologias para a descrição dos dados e processos de transformação e publicação de DAE.

- **Concepção do método proposto**: É o mapeamento do modelo do método. É constituído por um processo organizado em três fases (Extração, Transformação e Carga), com a finalidade de definir um metaprocesso (ou um *workflow*) de geração de dados abertos educacionais na abordagem ETL. O método leva em consideração três aspectos: fonte dos dados, processo de extração e principalmente a transformação dos dados submetidos ao longo do processo. Essa transformação inclui o enriquecimento semântico por meio do mapeamento dos dados públicos brutos para seus termos e ontologias, com o objetivo de agregar mais informações e assim criar um contexto ou significado ao dado. A transformação também inclui o processo de triplificação dos dados (sujeito, predicado e objeto) e a conversão para o formato RDF, planeando a publicação em um catálogo de dados abertos. Para a criação do diagrama, foi utilizada a ferramenta online diagrams.net (c2005-2021). Na figura 6, encontra-se a representação do método por fase:

FIGURA 6 – ARQUITETURA DO MÉTODO



Fonte: Elaborado pelo autor (2021)

- **Criação da aplicação**: Esta fase foi prevista para instanciar o método no contexto da UFMT e analisar a viabilidade prática do método com a utilização de um conjunto de dados reais para a produção e publicação de dados abertos no contexto da instituição. A aplicação *desktop* concebida foi implementada e orquestrada pela ferramenta ETL Kettle (versão pdi-ce-9.1.0.0-324 e versão 1.8.0_291 para o Java) juntamente com o *plugin* ETL4LOD+ (versão etl4lod 1.9). As funções da aplicação são três: extrair a

amostra de dados (levantados no Plano de Dados Abertos - PDA da UFMT) das bases da instituição; gerar indicadores educacionais em formato aberto (arquivos RDF serializados em xml) e disponibilizar os dados obtidos para publicação no catálogo CKAN (versão 2.9.4) da instituição; e o objetivo é a produção de dados abertos educacionais em formato RDF, enriquecidos mediante o uso de ontologias. O *download* do *plugin* foi realizado no repositório: https://github.com/johncurcio/ETL4LODPlus .

- **Verificação do método por grupo focal**: O grupo focal é composto por membros da área de Tecnologia da Informação e Comunicação, vinculados ao setor de TI de órgãos públicos (ambiente da pesquisa), selecionados de forma não probabilística, com representantes de IES públicas e do Governo do estado do Mato Grosso. Os participantes serão responsáveis por atestar a validade do método proposto, mediante o preenchimento de um questionário de pesquisa de opinião na escala Likert, que de acordo com Dalmoro e Vieira (2013) foi desenvolvida por Rensis Likert em 1932. O questionário virtual, dividido em 3 partes: perfil do participante, questões específicas sobre a opinião do participante e avaliação do grupo focal, será aplicado por meio da ferramenta Google Forms. O questionário referente à opinião do participante em relação ao método e aplicação é composto por cinco questões na escala Likert e uma em aberto para observações. As seguintes etapas foram seguidas: e-mail convite para um grupo de pessoas da área de TI através da ferramenta Google Calendar, reunião virtual via Google Meet para apresentação e conversa sobre o método e aplicação, aplicação do formulário, análise dos dados. Nos apêndices encontramos os documentos: O Termo de Consentimento Livre (A), questionário sobre o perfil (B), questionário avaliativo (C) e avaliação do grupo focal (D). A análise qualitativa dos dados da oficina será realizada a partir das transcrições das falas e com base nas opiniões recolhidas, melhorias serão realizadas no método. Para a análise quantitativa do formulário avaliativo será utilizado o Índice de Validade de Conteúdo (IVC) como método para quantificar o grau de concordância entre os especialistas (ALEXANDRE; COLUCI, 2011). O grupo focal será avaliado com uma questão na escala Likert e uma em aberto para comentários.

4 PROPOSTA DO MÉTODO

Para orientar a abordagem escolhida (ETL), o método proposto serviu-se da pesquisa bibliográfica sobre modelos, métodos, abordagens e ferramentas tecnológicas, encontrando nas respectivas soluções a inspiração necessária para produzir dados abertos governamentais. O método, composto por três fases de manipulação dos dados: extração (fase 1), transformação (fase 2) e carga (fase 3), busca resolver o problema de minerar os dados públicos da UFMT, transformando-os em dados abertos educacionais de qualidade no formato RDF, de forma a deixá-los prontos para publicação.

A abordagem ETL foi a escolhida levando-se em consideração os seguintes aspectos: facilidade e rapidez de desenvolvimento devido aos componentes visuais parametrizáveis (poupando a escrita de código), possibilidade de utilizar extensões ou *plugins* ampliando os recursos disponíveis para trabalhar com os dados durante o fluxo e disponibilidade de ferramentas de código aberto e gratuitas para uso.

4.1 Dados abertos no contexto da UFMT

A UFMT, na condição de IES pública mantida pelo poder federal, tem a obrigação legal de abrir os dados de acordo com a Lei de Acesso à Informação (LAI). Nesse sentido, conta com o Plano de Dados Abertos (PDA) que tem como objetivo geral (UFMT, 2021, n.p.):

> Promover a abertura de dados da Universidade Federal de Mato Grosso, zelando pelos princípios da publicidade, transparência e eficiência, visando o aumento da disseminação de dados e informações para a sociedade, bem como a melhoria da qualidade dos dados disponibilizados.

Estudando-se o PDA constatou-se que a UFMT não dispõe de um método automatizado para transformação de dados públicos brutos para dados abertos e que a promoção desta atividade compete às unidades responsáveis pela abertura dos dados. Segundo informado no PDA, a coleta e publicação destes dados deve ser prioritariamente automatizada, e alternativamente, semiautomatizada, cabendo ao STI (Secretaria de Tecnologia da Informação) disponibilizar um sistema para publicação automática dos dados (UFMT, 2021). Por oportunidade de inovação, esta proposta do método vem ao encontro das necessidades da UFMT para promover a abertura de dados, obedecendo a padrões mínimos de qualidade, de forma a facilitar o entendimento e a reutilização das informações.

A instituição também não possui ontologias oficinais para representação dos dados, com a autonomia dos setores para gerenciar os formatos dos documentos publicados. O plano de dados abertos orienta a publicar cada conjunto de dados com, no mínimo, os seguintes metadados (UFMT, 2021, n.p.):

a) Nome ou título do conjunto de dados;

b) Descrição sucinta;

c) Palavras-chave (etiquetas);

d) Assuntos relacionados do VCGE - Vocabulário Controlado do Governo Eletrônico;

e) Nome e e-mail do setor responsável pelos dados;

f) Periodicidade de atualização;

g) Escopo temporal (anual, mensal, diário, bimestral etc. exemplo: dados referentes ao censo de 2011, dados de um indicador mensal);

h) Escopo geopolítico (por cidade, por estado, por região).

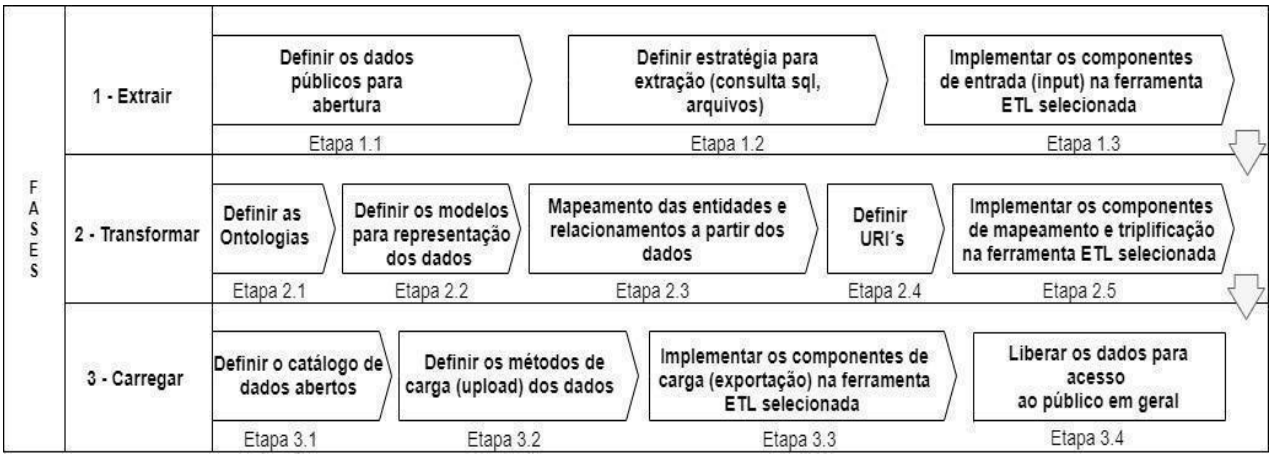
Após finalizada a publicação dos dados abertos institucionais no portal de dados abertos próprio - um catálogo de dados que utiliza a ferramenta CKAN, acessível a partir do endereço: https://dados.ufmt.br/, é gerada uma URL fixa que permite acesso direto ao conjunto de dados tanto por humanos quanto por agentes de *software*.

A partir do cenário institucional encontrado, um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT é proposto, visando a solucionar o problema por meio de uma abordagem ETL, conforme apresentado a seguir.

4.2 Método proposto

O método proposto são passos ou instruções para a mineração e transformação de dados abertos brutos em DAE, visando à aplicação do método em um processo de ETL (figura 7).

FIGURA 7 – ESQUEMA GERAL DO MÉTODO



Fonte: Elaborado pelo autor (2021)

A seguir, são apresentadas as três fases constitutivas do método, detalhadas por etapas, seguidas da respectiva e breve descrição:

Fase 1 - Extração dos dados: O método tem início com a extração dos dados brutos. Usualmente uma IES possui diversos sistemas (acadêmico, gerencial, recursos humanos, almoxarifado etc.) cujos dados estão distribuídos em diversas bases de dados, juntos com informações confidenciais.

Etapa 1.1: definição do conjunto de dados públicos para mineração e identificação das fontes de origem dos dados brutos. Após identificar as fontes, é necessário definir as estratégias de extração dos dados;

Etapa 1.2: definição da forma de extrair os dados de acordo com o tipo da fonte de dados (exemplos: banco de dados, planilha, documento de texto). É possível definir uma pasta como repositório dos arquivos ou criar consultas SQL, no caso de um banco de dados relacional.

Etapa 1.3: implementação dos componentes de entrada (*input*) na ferramenta ETL selecionada. As ferramentas possuem uma variedade de componentes que possibilitam ler os dados de arquivos xml, csv, banco de dados, entre outros, conforme a realidade de cada instituição. Com a abordagem ETL é possível modularizar as fases, aumentando seu potencial de reuso em outras universidades e adaptando o método para o contexto de cada órgão público.

Fase 2 - Transformação dos dados: Antes de realizar o mapeamento dos dados brutos para seus termos é necessário definir quais ontologias serão utilizadas para representar os dados.

Etapa 2.1: definição das ontologias para a anotação semântica dos dados originais.

Etapa 2.2: uso do modelo em RDF para exprimir a ideia da definição dos dados por meio de um conjunto de triplas, com o sujeito (relação chamada predicado) e o nó objeto (sujeito, predicado, objeto).

Etapa 2.3: estabelecimento da correspondência entre campos vindos do banco de dados ou arquivos para seus respectivos termos e vocabulários previamente escolhidos.

Etapa 2.4: estabelecimento do mecanismo de identificação e do acesso único aos conjuntos de dados, o padrão de URI a ser adotado pela organização. Uma forma de permitir acesso aos conjuntos de dados é o uso de URIs em http (uma url do protocolo http).

Etapa 2.5: uso dos recursos dos componentes de transformação da ferramenta de ETL para enriquecer os dados semanticamente com os modelos e ontologias escolhidas. Outro passo importante desta etapa é a serialização dos modelos RDFs em algum formato suportado

(JSON-LD, RDFa -código RDF embutido em HTML-, RDF/XML, NTriples). Exemplificando, um modelo RDF pode ser serializado em um arquivo no formato html e seus atributos representados com as tags html, permitindo o armazenamento das triplas em arquivo texto.

Fase 3 - Carga dos dados: Após o enriquecimento semântico e a transformação dos dados, passamos a ter DAE prontos para a publicação. Para liberar os dados para consultas públicas é necessário carregar (ou fazer o *upload*) para o canal de comunicação escolhido.

Etapa 3.1: estabelecimento do canal (qual a forma de comunicação a ser usada com os consumidores, podendo ser o *site* da instituição, *webservices* ou uma ferramenta de catalogação de dados).

Etapa 3.2: determinação do procedimento de carga do conjunto de dados que depende da ferramenta de catalogação definida. A carga pode ser via API, inserção direta em banco de dados ou exportação de arquivos, tudo depende da infraestrutura selecionada.

Etapa 3.3: implementação da carga nos componentes de saída (*output*) da ferramenta ETL, permitindo salvar ou carregar os dados serializados em arquivos json, csv, excel, xml, banco de dados, entre outros.

Etapa 3.4: disponibilização dos DAEs para consulta pública.

O quadro 2 resume o método, mostrando as informações e o esperado para cada fase do processo, consoante aspectos técnicos (como os recursos tecnológicos de cada fase) indispensáveis à aplicação *desktop* em ETL, a ser apresentada na seção 4.3:

QUADRO 2 – RESUMO DO MÉTODO

Fases		Objetivo	Descrição da fase	Entrada	Saída	Aspectos técnicos	Ferramenta
1	Extract	Minerar os dados brutos de domínio públicos no banco de dados, arquivos ou nas nuvens (API)	Na fase de extração os dados públicos brutos vindos de diversas fontes possíveis da instituição serão extraídos para posterior uso e enriquecimento	API´s, BD, Arquivos	Conjunto de dados público brutos	Com a abordagem e as ferramentas ETL é possível extrair dados de um banco de dados relacional com uma consulta SQL ou ler os dados de arquivos ou APIs (json)	Kettle
2	Transform	Enriquecer os dados semanticamente	Os dados brutos devem ser mapeados com base em modelos de referências, ontologias e metadados, agregando valor semântico	Conjunto de dados público brutos	Arquivos RDF enriquecidos semanticamente	A transformação dos dados é realizada pelo plugin ETL4LOD+ do Kettle. Nesta fase o modelo para representação dos dados será aplicado aos dados brutos, juntamente com a triplicação (arquivos RDF)	Plugin ETL4LOD+
3	Load	Gerar dados abertos governamentais prontos para publicação	Nesta fase os dados estão prontos (em formato RDF) para a publicação em um catálogo de dados abertos	Grafo RDF	Dados disponíveis para exportação (e consulta) no catálogo da instituição	Os dados estão prontos para serem carregados no catálogo de dados abertos da instituição através das chamadas aos serviços (API) de inserção do catálogo. Após carregados estão prontos para consulta pública	CKAN

Fonte: Elaborado pelo autor (2021)

4.3 Utilização do método: aplicação *Desktop* em ETL

Esta seção detalha a demonstração do método com um exemplo real, com a abertura de uma amostra de dados da UFMT, fazendo uso da abordagem proposta pelo método. Para

testar a viabilidade do método, uma versão funcional de uma aplicação será implementada com o Pentaho Data Integration ou Kettle. Duas são as grandes vantagens das ferramentas de ETL: a programação visual e os componentes parametrizáveis representados por ícones, ou seja, o uso de uma interface gráfica para a criação de aplicações, aliado ao conceito de fluxo de trabalho, com os dados percorrendo um curso da entrada (componentes de *input*) até a saída (*output*), passando por transformações durante o processo.

A disponibilização de uma *interface* gráfica no lugar da programação por linhas de código e os componentes ajustáveis permitem uma facilidade na aplicabilidade da solução para a realidade de outras instituições, alterando apenas alguns parâmetros ou componentes pontuais. Santos (2016, p. 32) apresenta o seguinte conceito sobre a ferramenta *Spoon* do Kettle: "ferramenta de *interface* gráfica para desenvolvimento e execução do fluxo de dados, desde sua entrada até a saída, através da criação e execução de *jobs* e *tranformations*". No Kettle, o processo ETL pode ser especificado com o uso de dois tipos de recursos chamados *Jobs* e *Transformations*. Eles determinam a ordem do processo geral e do fluxo de dados. Santos (2016, p. 31) define *job* e *transformation* da seguinte maneira:

> Um *transformation* consiste de um conjunto de passos conectados, onde cada passo, denominado *step*, é responsável por uma atividade de extração, transformação ou carga de dados.
>
> Um *job* também consiste de um conjunto de passos conectados. No entanto, os passos de um *job*, denominados *job entries*, são responsáveis por executar um *transformation*, outro *job* ou atividades auxiliares como manipular e transferir arquivos, enviar e receber emails e executar uma série de validações.

O centro da solução proposta nesta seção é o *plugin* ETL4LOD+ para o Kettle. O ETL4LOD+ é um *framework* filho (ou uma extensão) do ETL4LOD, desenvolvido pelo Grupo de Engenharia do Conhecimento - GRECO da UFRJ e fazia parte de uma plataforma maior denominada LinkedDataBR, que tem por objetivo propor novas soluções para limpeza e transformação (triplificação) de dados no contexto de dados abertos conectados (SILVA, 2018). O *plugin* permite o tratamento, a triplificação e a publicação de dados conectados (SILVA, 2018) que contenham recursos para a geração de triplas RDF (SILVEIRA, 2021). A extensão ETL4LOD possui os seguintes componentes para executar as atividades relacionadas a dados conectados (SILVEIRA, 2021, p. 67; SILVA, 2018, p. 32):

> - **Data Property Mapping**: fornece a habilidade de mapear os campos de entrada em triplas RDF, indicando quem é o sujeito, objeto e predicado da tripla quando o objeto é um literal. Também permite que esse literal seja anotado com informações do tipo de dado e da linguagem usada.
> - **Object Property Mapping**: responsável pela transformação dos dados de entrada em triplas RDF. Permite a anotação dos relacionamentos existentes entre os dados de um determinado domínio a partir dos conceitos presentes em uma ontologia de referência. Possui um funcionamento similar ao do Data Property Mapping, porém os objetos mapeados no Object Property Mapping são URIs ao invés de literais.

- **NTriple Generator**: Gera triplas RDF no formato N-Triples.
- **Sparql Endpoint**: permite a construção de consultas SPARQL (do tipo SELECT) em endpoints específicos através do processo de publicação.
- **Sparql Run Query**: permite executar uma consulta para atualizar um conjunto de dados, tal como UPDATE, DELETE e DROP. Essa consulta precisa estar definida em um campo vindo de um plug-in anterior e cada linha desse campo precisa ser uma consulta inteira a ser executada.

Agora que algumas características da ferramenta ETL foram apresentadas, tem início a explicação do método na prática, com o desenvolvimento da aplicação, conforme as etapas propostas:

- **Etapa 1.1**: Vamos considerar como amostra de dados a lista de dados levantados para abertura do anexo 1 do plano de dados abertos da UFMT. A lista contém itens como o quantitativo de estudantes bolsistas, estudantes em mobilidade, estudantes estrangeiros ingressantes por políticas afirmativas na pós-graduação e estudantes formados na graduação (UFMT, 2021). O resultado da etapa 1.1 encontra-se demonstrado no quadro 3.

QUADRO 3 – CONJUNTOS DE DADOS LEVANTADOS

Número da Amostra	**Descrição dos dados**	**Localização**	**Armazenamento**
1	Quantitativo de estudantes da pós-graduação em mobilidade	SIPG	Banco de dados relacional
2	Quantitativo de estudantes estrangeiros da pós-graduação	SIPG	Banco de dados relacional
3	Quantitativo de estudantes da pós-graduação	SIPG	Banco de dados relacional
4	Quantitativo de estudantes da graduação	SIGA	Banco de dados relacional
5	Quantidade de estudantes graduação formados	SIGA	Banco de dados relacional

Fonte: Elaborado pelo autor (2021)

- **Etapa 1.2**: Os conjuntos de dados selecionados no presente contexto estão armazenados em um banco de dados relacional. A estratégia de extração consiste em

consultas SQL (SQL *queries*). Esta etapa é dependente da realidade de cada instituição e as consultas levam em consideração características técnicas particulares como o nome das tabelas e das colunas do banco de dados ou do servidor de banco. Exemplo de uma consulta SQL (*select * from ViewMatriculadosPeriodo where periodo=20211*) cujo resultado está disposto no quadro 4.

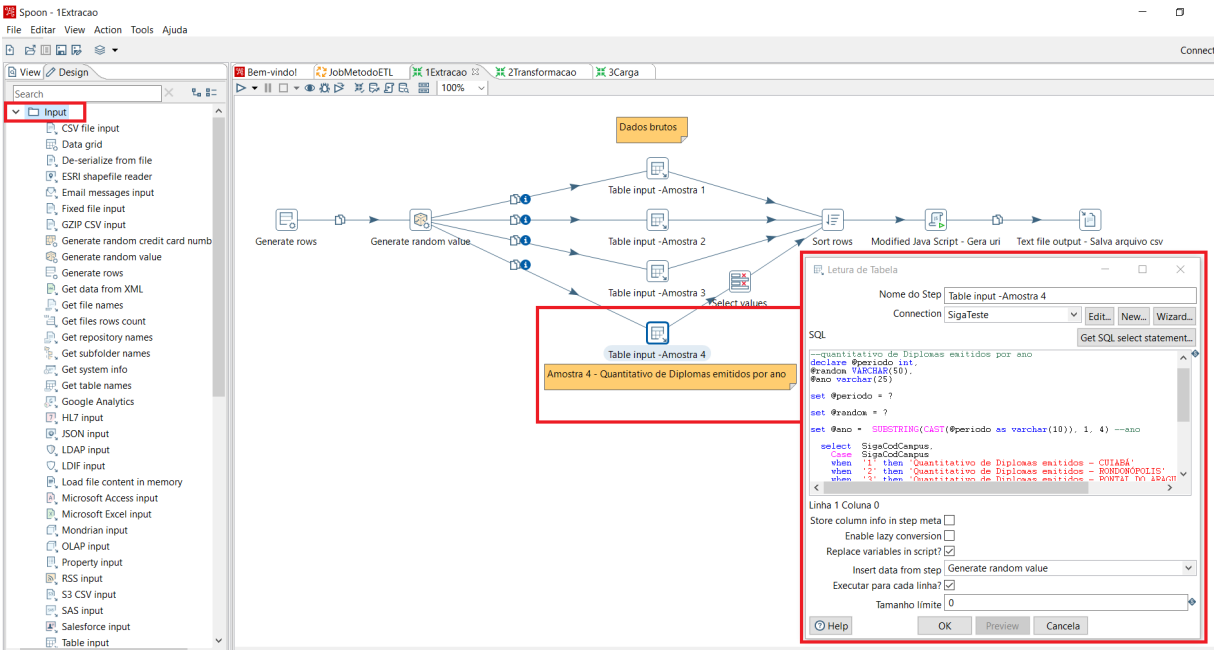
QUADRO 4 – RESULTADO COM O QUANTITATIVO POR CURSO

Período	**Curso**	**Total**
20211	PROGRAMA DE PÓS-GRADUAÇÃO EM ZOOTECNIA	20
20211	PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA	31
20211	PROGRAMA DE PÓS-GRADUAÇÃO EM HISTÓRIA	27
20211	PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA	15

Fonte: Elaborado pelo autor (2021)

- **Etapa 1.3**: Após identificadas as fontes e os dados selecionados, o próximo passo é implementar os componentes de entrada da ferramenta ETL. Na figura 8, encontram-se o *transformation* de extração e os componentes de *input* (com destaque para o *Table Input* com a consulta SQL).

FIGURA 8 – EXTRAÇÃO DOS DADOS NO KETTLE



Fonte: Elaborado pelo autor (2021)

- **Etapa 2.1**: Nesta fase 2, inicia o processo de enriquecimento e transformação referente às escolhas de ontologias a serem utilizadas para representar os dados, como demonstrado no quadro 5.

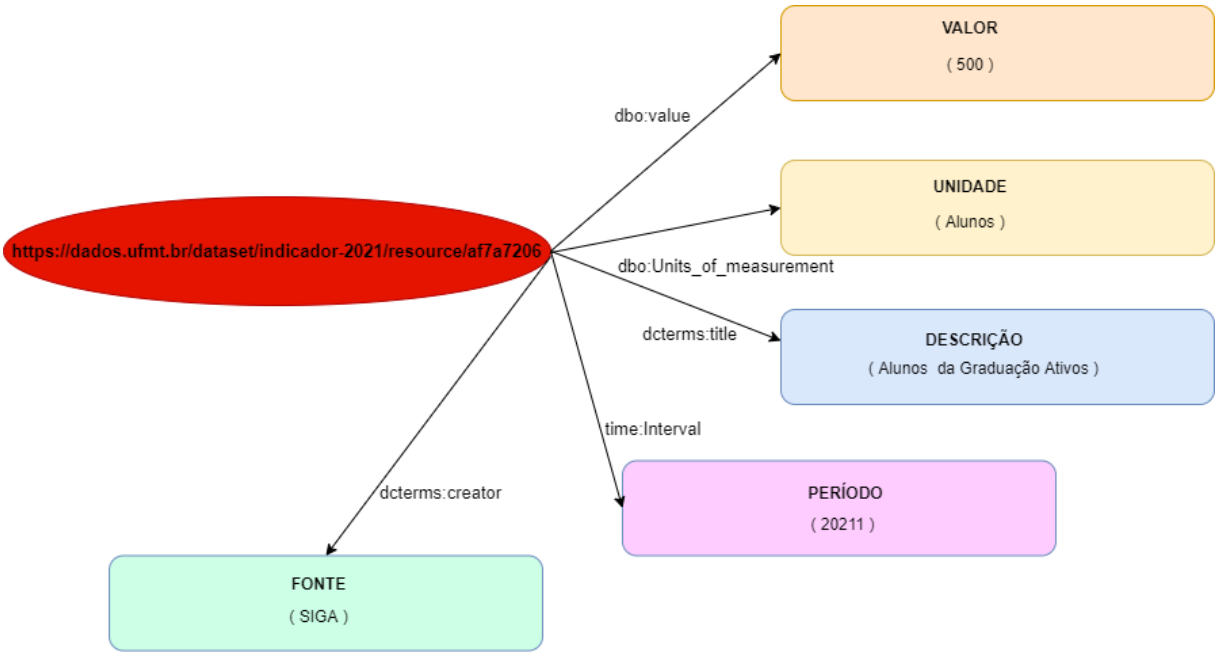
QUADRO 5 – ONTOLOGIAS

Ontologia	Descrição	Prefixo
Dbpedia	Ontologia usada como fonte de dados de definição de classes	dbo:<http://dbpedia.org/ontology/>
Dublin Core	Ontologia adotada para as propriedades (descreve artefatos digitais)	dcterms:<http://purl.org/dc/terms/>
Time Ontology	Ontologia para descrever intervalos de tempo	time:<https://www.w3.org/TR/owl-time/>

Fonte: Elaborado pelo autor (2021)

- **Etapa 2.2**: Nesta etapa são modelados os dados conforme os modelos de representação determinados (ver figura 9).

FIGURA 9 – MODELO PARA REPRESENTAÇÃO DOS DADOS



Fonte: Elaborado pelo autor (2021)

- **Etapa 2.3**: O mapeamento consiste em fazer as correspondências entre os campos vindos do banco de dados (e da fase de extração) e os termos correlatos das ontologias. O quadro 6 apresenta o resultado da etapa.

QUADRO 6 – MAPEAMENTO ENTRE CAMPOS E TERMOS

Campo do arquivo	Propriedade (Vocabulário)	Tipo do Dado
fonte	dcterms:creator	String
período	time:Interval	Integer
local	dbo:locationOf	String
indicador	rdf:type	Recurso (ou sujeito)
valor	dbo:value	Integer
unidade	dbo:Units_of_measurement	String
descrição	dcterms:title	String

Fonte: Elaborado pelo autor (2021)

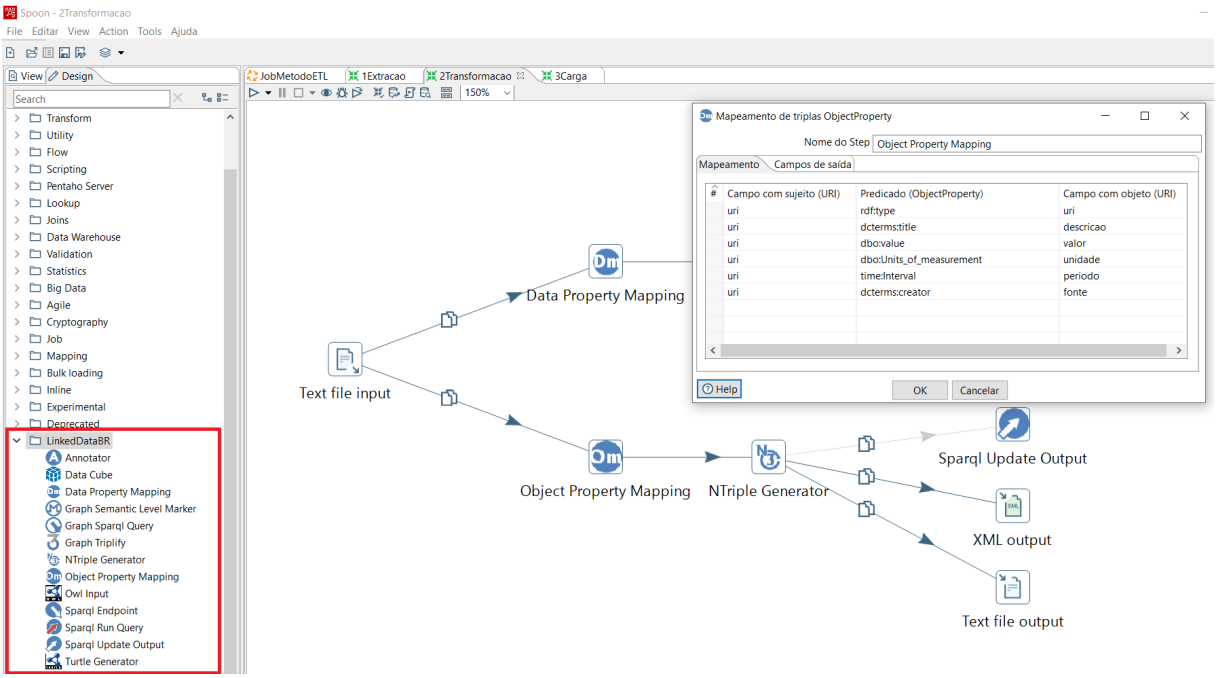
- **Etapa 2.4**: Para definição de URI, foi adotado o padrão utilizado no portal de dados aberto da UFMT, que consiste em um catálogo de dados com CKAN, com a ferramenta criando uma URL única para cada conjunto de dados publicados. Vejamos este exemplo de URL gerada pelo CKAN (figura 10):

FIGURA 10 – URL GERADA PELO CKAN

Fonte: Elaborado pelo autor (2021)

- **Etapa 2.5**: Nesta etapa são utilizados os componentes do *plugin* ETL4LOD para mapear (*Data e Object Property Mapping*) e converter os dados para RDF (*NTriple Generator*). O *plugin* permite o enriquecimento semântico com o uso das ontologias e a triplificação do modelo para RDF. A figura 11 apresenta os componentes de transformação.

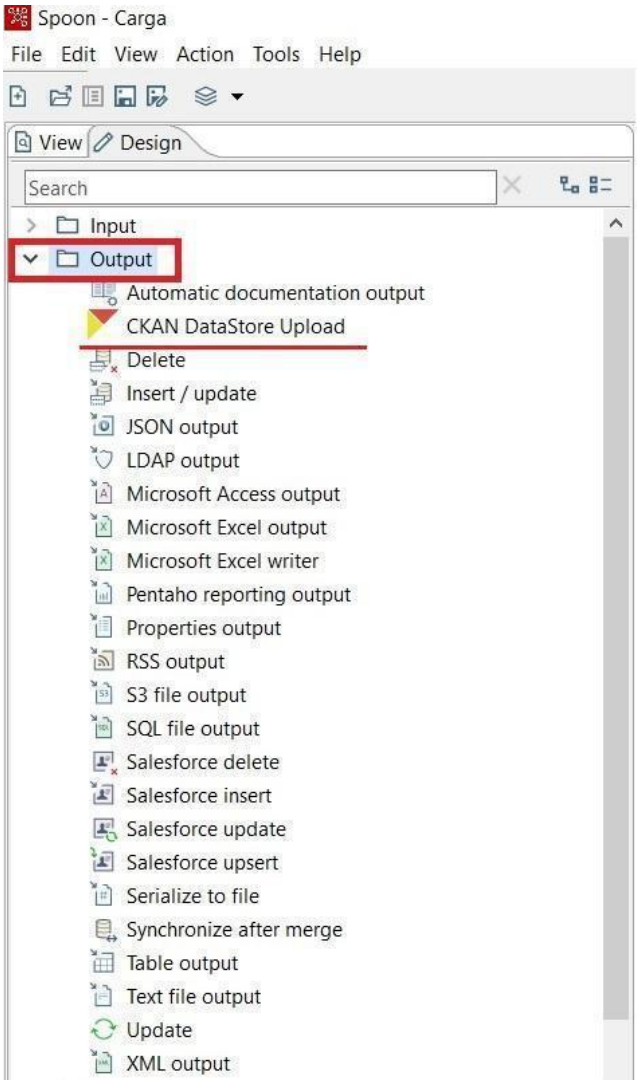
FIGURA 11 – COMPONENTES DO ETL4LOD



Fonte: Elaborado pelo autor (2021)

- **Etapa 3.1**: Definir o catálogo de dados abertos para que os DAEs gerados nas etapas anteriores sejam publicados. A publicação depende da infraestrutura tecnológica da instituição e a UFMT adotou o CKAN como ferramenta de catalogação.
- **Etapa 3.2**: Para a carga dos dados, utilizam-se os recursos de saída (*output*) do Kettle. Entre as possibilidades de saída da ferramenta temos arquivos JSON, CSV, XML etc ou inserção direta no banco de dados. De uma série de opções, foi adotado o componente CKAN *DataStore Upload*:

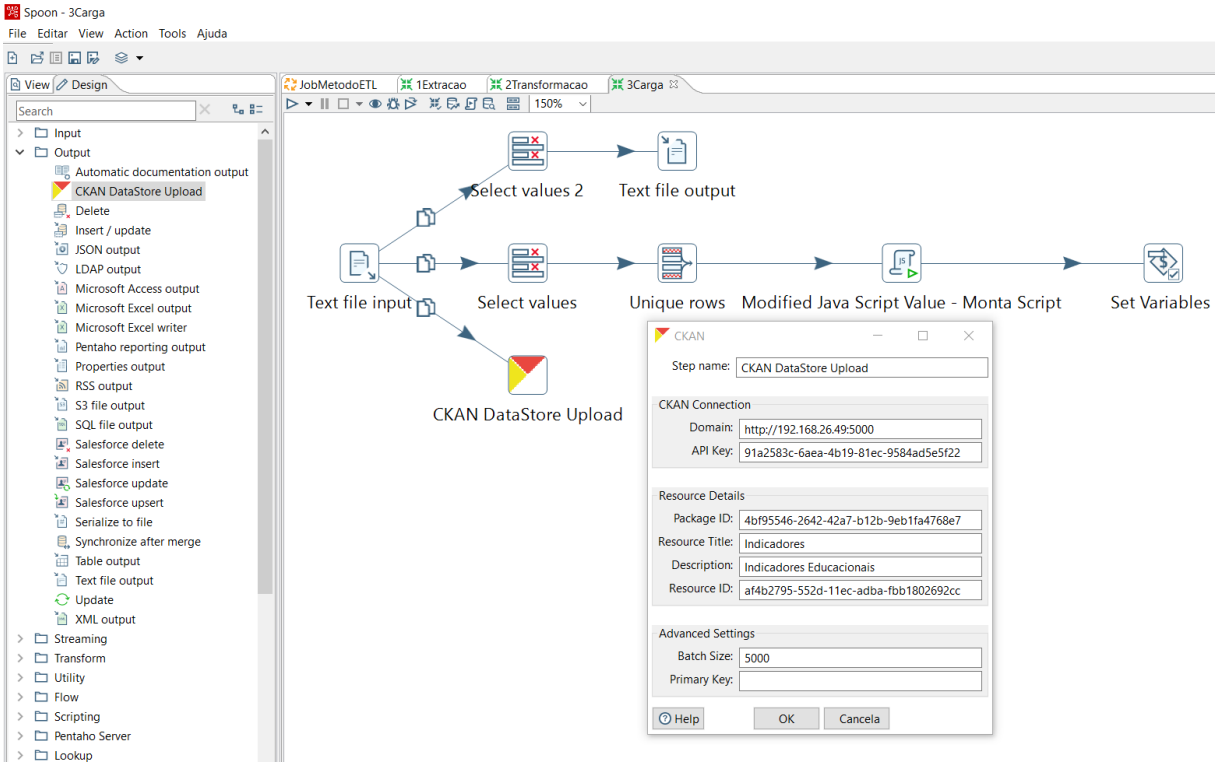
FIGURA 12 – COMPONENTE DE SAÍDA/CARGA



Fonte: Elaborado pelo autor (2021)

- **Etapa 3.3**: O componente de saída CKAN *DataStore Upload* permite fazer *upload* dos dados direto para o catálogo via API e, com a alteração dos parâmetros, torna fácil a adaptação da saída para outros catálogos hospedados em servidores diferentes. A figura 13 apresenta a configuração do componente.

FIGURA 13 – CKAN DATASTORE UPLOAD

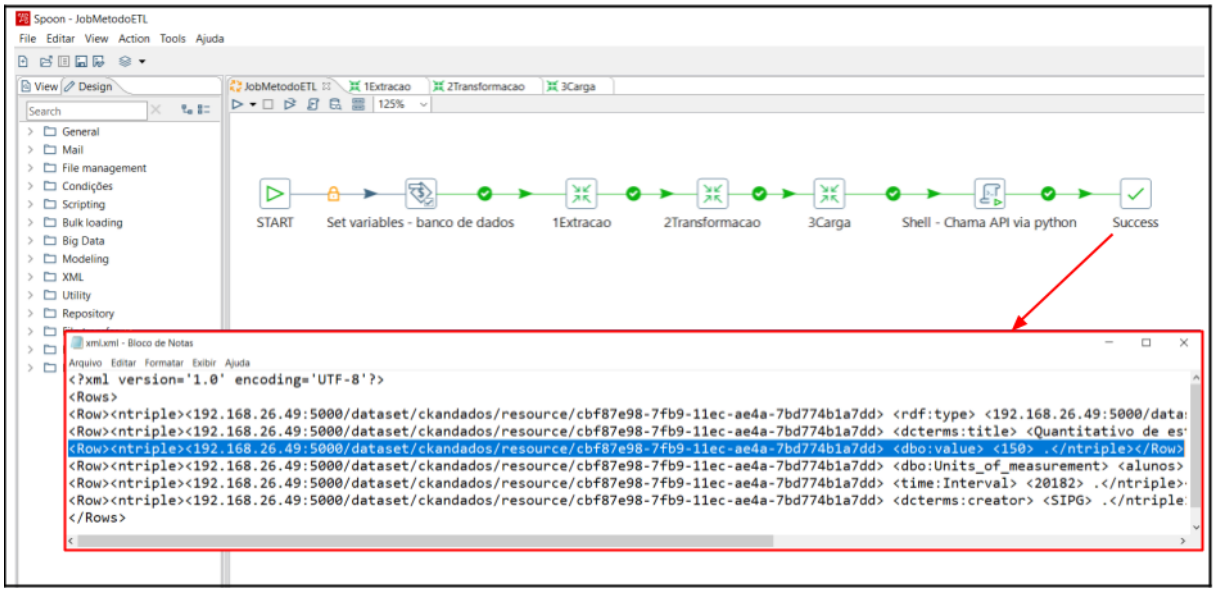


Fonte: Elaborado pelo autor (2021)

- **Etapa 3.4**: Após a carga no catálogo, os dados abertos estão prontos para consulta pública no portal da instituição.

Este capítulo descreveu o desenvolvimento da aplicação, demonstrando os passos seguidos no método proposto. As ferramentas de ETL disponíveis na atualidade oferecem um leque de opções e meios para extrair e transformar dados. Com poucas alterações nos componentes e em seus parâmetros, é possível expandir ou moldar a solução para o cenário de cada instituição, a exemplo de aspectos técnicos, como a realidade dos servidores e da infraestrutura e os referentes às ontologias e termos de cada universidade. Existe também um ganho de tempo na construção da solução proporcionados pelos componentes gráficos do Kettle, dispensando a escrita de códigos de programação e lógicas próprias. O *workflow* completo e implementado no Pentaho Data Integration é apresentado na figura 14, com ênfase no arquivo xml de saída com as triplas RDF.

FIGURA 14 – *WORKFLOW* ETL



Fonte: Elaborado pelo autor (2021)

5 ANÁLISE DOS RESULTADOS

Conforme exposto ao longo do presente trabalho, a publicação automatizada de dados abertos governamentais representa uma contribuição de valor para a transparência do setor público. Esta seção versa sobre a aplicabilidade do método e os resultados obtidos com a aplicação *desktop* ETL, dada a realização de uma oficina com o grupo focal no dia 07/12/2021, às 14h.

A execução da oficina ocorreu de forma remota através de uma videochamada na ferramenta Google Meet. O autor desta dissertação foi o condutor da oficina, sendo o orientador um observador, o qual só fez interferências pontuais e agradeceu aos presentes. Os trabalhos iniciaram com a apresentação do método na forma de *slides*, sucedendo com a demonstração da aplicação com um exemplo real. Após as demonstrações, foi aberto um espaço para conversa, finalizando a experiência com o preenchimento dos formulários via Google Forms.

5.1 Perfil dos participantes

A proposta foi avaliada por um grupo de 8 pessoas da área de TI de instituições públicas após a apresentação do autor e a conversação. O grupo focal reuniu 1 doutor(a), 2 mestres, 4 especialistas e 1 graduado(a), todos com atuação profissional no setor público.

Seguindo com a análise do perfil dos membros do grupo focal, a partir da resposta referente ao grau de conhecimento sobre dados abertos podemos constatar a homogeneidade dos participantes. Do total de 8 pessoas, 50% consideraram seus conhecimentos sobre dados abertos como básico, nenhum ou pouco. Enquanto os outros 50% avaliaram seus conhecimentos como bom, avançado ou conceitual.

5.2 Análise dos resultados do grupo focal e do formulário de avaliação

Após a apresentação da solução pelo autor, a verificação do método ocorreu de duas formas: um momento aberto para uma conversa sobre o método e a solução e o formulário referente a opinião dos participantes.

A conversação teve como objetivo verificar a validade e a aplicabilidade do método proposto e capturar possíveis pontos que pudessem gerar dúvidas no trabalho. Nas observações levantadas no momento do debate foram destacados diversos benefícios e elogios a solução, as seguintes contribuições foram ressaltadas:

> Comentário 1: Uma das bases da viabilidade do trabalho do Fabio é bem interessante, é justamente nessa questão de alimentação mais automática do ckan na plataforma de dados abertos.
>
> Comentário 2: A questão desse passo a mais que ele fez, que foi a triplicação dessas informações, já se preocupando com essa necessidade, muitas vezes, muitos trabalhos de dados abertos o pessoal só faz a extração de um lugar, transforma esses dados e carrega em outro mas não chega a trabalhar toda essa semântica, a ontologia por trás dessas informações, eu achei muito legal.
>
> Comentário 3: Com essa metodologia isso hoje é totalmente aplicável , a gente não precisa necessariamente colocar uma funcionalidade em cada um dos sistemas ou não necessariamente precisa resgatar o conhecimento daquela sistema, pode ir direto na fonte dos dados , fazer a pergunta para fonte de dados de qual é a informação, gera essa consulta e dá a publicidade devida de maneira muito mais ágil, então a gente tem uma ferramenta extremamente interessante que vai encaixar muito bem na instituição.

Os comentários acima enfatizaram as características da automatização do processo e a triplificação dos dados, dois fatores de extrema importância para o alcance dos objetivos do trabalho, demonstrando a importância da solução para a UFMT.

A análise dos dados coletados no grupo focal também justificou melhorias no método proposto, incorporando como melhoria ao trabalho a figura 7, com a representação mais detalhada do método, dividindo-o em etapas. A melhoria partiu do seguinte comentário de um dos participantes do grupo focal:

> Comentário 4: Aí fica a sugestão de pensar talvez isso, em colocar numa BPMN o processo, a título de melhoria de entendimento do processo de construção que vocês estão propondo (...) BPMN fosse interessante nesse sentido porque uma das perguntas que vocês colocam é até isso do entendimento do método.

Continuando com a análise, o formulário referente à opinião do participante em relação ao método e aplicação é composto por cinco questões na escala Likert e uma em aberto e os resultados das análises destas questões avaliativas serão comentados nesta seção.

Este trabalho empregou uma escala do tipo Likert com pontuação de um a cinco (5 = Concordo Totalmente, 4 = Concordo Parcialmente, 3 = Não Concordo Nem Discordo, 2 = Discordo Parcialmente, 1 = Discordo Totalmente) para avaliar a relevância/concordância das respostas e para a análise quantitativa do formulário sobre o sentimento/opinião do participante em relação a solução optou-se pelo cálculo do Índice de Validade de Conteúdo (IVC). A fórmula utilizada foi IVC = número de respostas 4 ou 5 dividido pelo número total de respostas (IVC = (número de respostas 4 ou 5) / (número total de respostas)). A porcentagem de concordância e o índice de validade de conteúdo (IVC) referente às questões específicas sobre a opinião do participante são:

Questão 1 - Houve 100% de concordância com o IVC de 1,00 para a afirmação que “A solução está de acordo com o PDA da UFMT no quesito de propor uma forma automatizada

para publicação dos dados abertos, com ganhos de eficiência em comparação a extrações pontuais.”.

Questão 2 - Houve 100% de concordância com o IVC de 1,00 para a afirmação que “O método/aplicação são escaláveis para outros conjuntos de dados ou instituições”.

Questão 3 - Houve 100% de concordância com o IVC de 1,00 para a afirmação que “O método com foco na abordagem ETL e a ferramenta *open source* Kettle atendem os requisitos tecnológicos para geração de Dados Abertos Governamentais”.

Questão 4 - Houve 87,50 % de concordância com o IVC de 0,875 para a afirmação que “A aplicação e o método são de simples compreensão”.

Questão 5 - Houve 100% de concordância com o IVC de 1,00 para a afirmação que “A solução atendeu ao objetivo de propor um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT”.

A questão 5 continha um campo em aberto para comentário, solicitando a justificativa da resposta e foram feitas as seguintes observações:

Participante 1 - “A ferramenta permite facilmente que o setor de TI de qualquer órgão implemente a importação de dados para a carga e para dar a publicidade necessária aos dados”.

Participante 2 - "Método de fácil entendimento”.

Participante 3 - “Muita boa a solução de dados abertos”.

Participante 4 - “Aproveitando para reforçar os pontos que mencionei, que foram a possibilidade (futura ou não) de adicionar uma notação como a BPMN para tornar o método mais claro e até mesmo poder automatizá-lo de alguma forma; e o segundo comentário, pensar nas questões de start do processo a partir de demandas dos cidadãos provenientes da LAI, e não somente de demandas internas”.

Participante 5 - “Pois ele faz o uso da ferramenta CKAN e utiliza dele o método de publicação dos dados”.

Participante 6 - “O método com utilização de ETL, se torna muito flexível e acessível para qualquer instituição ou fonte dos dados. Porém, quanto mais heterogênea forem as fontes de dados, mais complexo o processo pode se tornar, dificultando um pouco a compreensão de todo o processo para pessoas fora da área de análise de dados / Banco de Dados”.

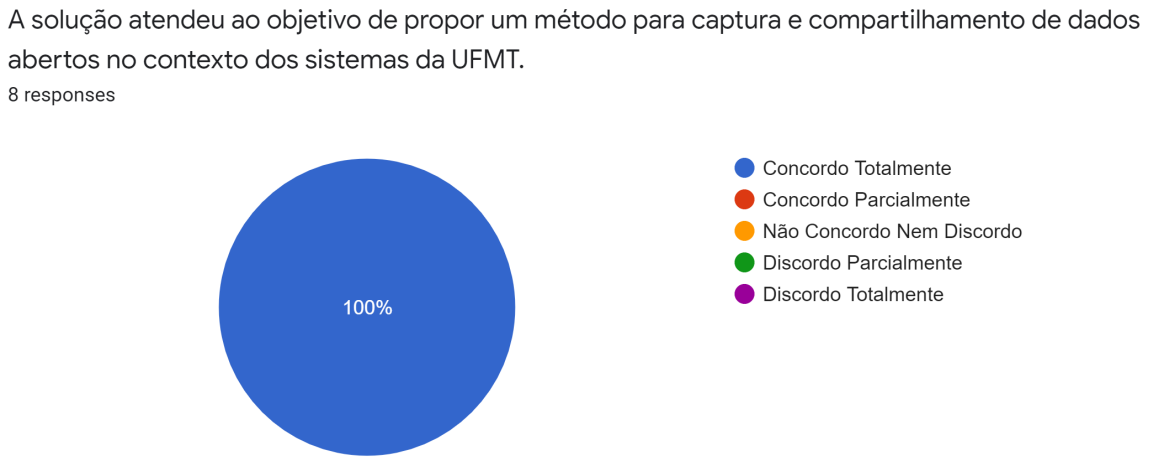
Participante 7 - “A solução é uma ferramenta de extrema importância para sociedade, pois auxilia a ação de concentrar, coletar e compartilhar os dados em um ambiente web de forma prática e de fácil acesso”.

Participante 8 - “A importância da disponibilização dos dados abertos permite a participação social e a necessidade por melhores serviços públicos são temas de ordem, a política de abertura de dados e é insumo fundamental para a construção e a consolidação do governo aberto e inteligente. Dessa forma permite ao cidadão obter informações sobre as ações da instituição/governo, tornando possível sua contribuição ativa no processo de decisão e melhoria do funcionamento do Estado”.

Desta forma, após a realização do grupo focal considerou-se (fundamentado nas avaliações dos participantes da oficina) que o método e a solução atenderam ao objetivo de propor um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT. As afirmações dos participantes destacadas na sequência demonstram a aceitação do método: Participante 1: "A ferramenta permite facilmente que o setor de TI de qualquer órgão implemente a importação de dados para a carga e para dar a publicidade necessária aos dados", Participante 2: "método de fácil entendimento", Participante 7: "A solução é uma ferramenta de extrema importância para sociedade, pois auxilia a ação de concentrar, coletar e compartilhar os dados em um ambiente web de forma prática e de fácil acesso".

A figura 15 apresenta o gráfico com as respostas da questão 5, confirmando que a solução atendeu ao objetivo proposto segundo os participantes do grupo focal:

FIGURA 15 – GRÁFICO COM AS RESPOSTAS DA QUESTÃO 5



Fonte: Elaborado pelo autor (2021)

Sobre uma taxa de concordância aceitável para o IVC, Alexandre e Coluci (2011) consideram:

> Com a participação de cinco ou menos sujeitos, todos devem concordar para ser representativo. No caso de seis ou mais, recomenda-se uma taxa não inferior a 0,78. Para verificar a validade de novos instrumentos de uma forma geral, alguns autores sugerem uma concordância mínima de 0,80. No entanto, neste caso os valores recomendados devem ser de 0,90 ou mais.

O resultado do IVC médio das cinco questões específicas sobre o sentimento dos 8 participantes em relação à solução proposta foi de 0,975, obtendo um nível muito bom de concordância/aceitação. O cálculo do IVC é apresentado na tabela 1:

Tabela 1 – Cálculo do IVC médio

	Participante 1	Participante 2	Participante 3	Participante 4	Participante 5	Participante 6	Participante 7	Participante 8	IVC
Questão 1	5	5	5	5	5	5	5	5	8/8 = 1
Questão 2	5	5	4	5	5	5	5	5	8/8 = 1
Questão 3	5	5	4	5	5	5	5	5	8/8 = 1
Questão 4	4	5	5	3	4	4	4	5	7/8 = 0,875
Questão 5	5	5	5	5	5	5	5	5	8/8 = 1
								Média:	0,975

Fonte: Elaborado pelo autor (2021)

Após a verificação do método quanto ao seu objetivo e conteúdo, a dinâmica com o grupo focal foi avaliada pelos 8 participantes com a pergunta “Como você avalia a experiência do grupo focal?”, resultando em 5 respostas “Muita boa”.

Finalizando a análise do grupo focal, o comentário positivo de um dos participantes sobre a dinâmica da oficina encerra de forma satisfatória a experiência:

> A dinâmica do grupo foi muito interessante, pois os participantes trouxeram as suas experiências do cotidiano diante da ausência e/ou dificuldade de se obter informações. Foi comprovado que a ferramenta apresentada é um facilitador e será de grande utilidade para quem precisa divulgar e dar mais transparência às informações. O palestrante se mostrou seguro, confiante sobre o assunto e comprovou que seu projeto irá contribuir muito para a concentração e compartilhamento de dados.

5.3 Demonstração do método

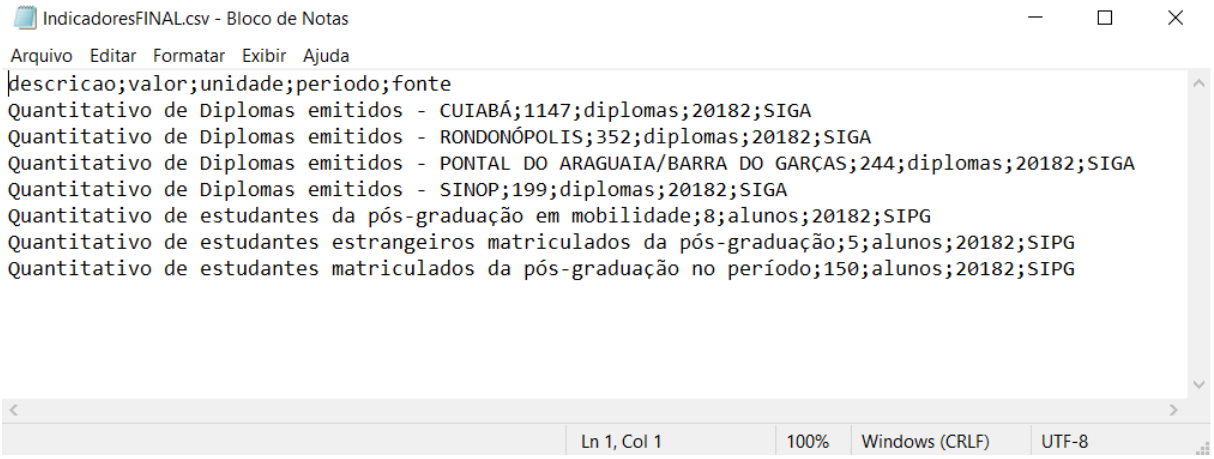
A proposta de um método genérico permite a execução das etapas em diferentes contextos de infraestrutura tecnológica ou fontes de dados e a implementação em diversas ferramentas de ETL disponíveis para uso na atualidade. O presente trabalho implementou as etapas do método na ferramenta Kettle juntamente com o *plugin* ETL4LOD+, focando em

propor um método aplicável na realidade da UFMT, atendendo a uma demanda interna da universidade, porém escalável para outras unidades ou instituições.

Para demonstração do método uma aplicação ETL com as funcionalidades necessárias para cumprir ao objetivo de capturar e compartilhar dados abertos no contexto dos sistemas da UFMT foi implementada. Percorrendo as etapas do método proposto foi possível criar uma solução funcional, com um conjunto de dados reais: indicadores educacionais da universidade, finalizando com a publicação no catálogo CKAN. Após executar o *workflow* proposto os resultados obtidos em cada etapa e a transformação dos dados durante o processo, com ênfase nas entradas e saídas de cada fase, são o foco das análises a seguir:

Fase 1 - Extração dos dados: o objetivo desta fase foi minerar os dados brutos do banco de dados relacional da universidade. Após a definição da amostra de dados e as consultas SQLs implementadas nos componentes de *input* do Kettle (conforme apresentado na Figura 8 - Extração dos dados no Kettle) o resultado da fase 1 ao final da etapa 1.3 foi um arquivo no formato csv (formato não proprietário, legível por máquina e de licença aberta). O arquivo .csv gerado pelo Kettle ao final da extração (fase 1) é apresentado na figura 16:

FIGURA 16 – ARQUIVO EM FORMATO CSV GERADO AO FINAL DA FASE 1

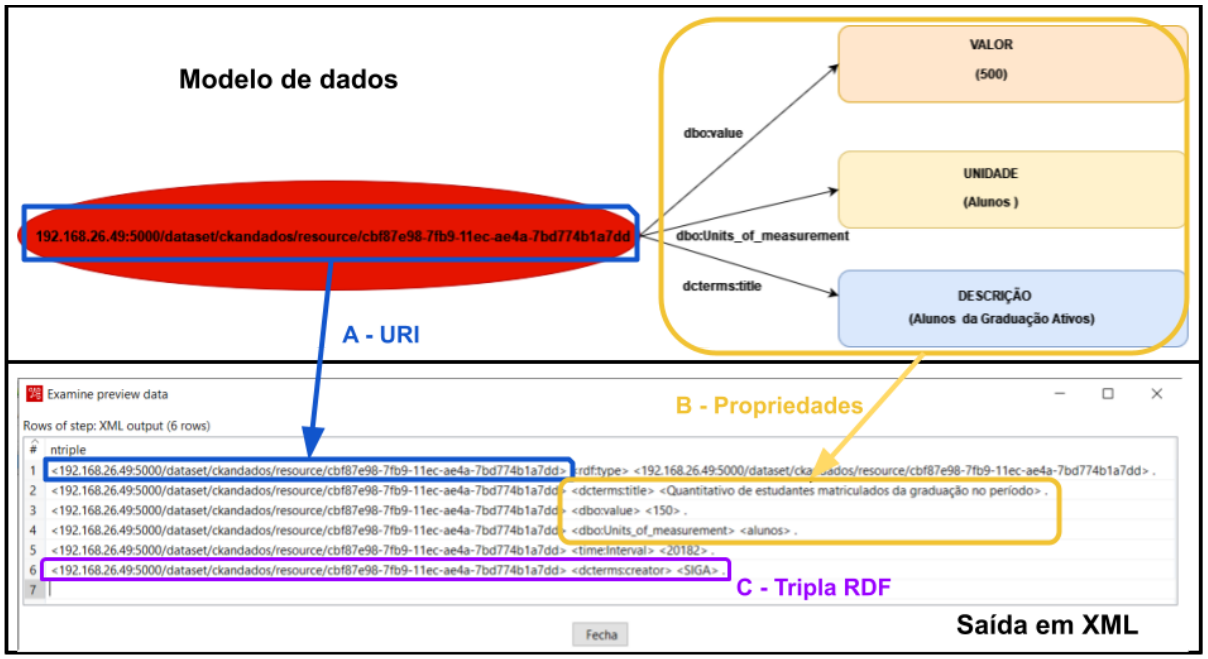


Fonte: Elaborado pelo autor (2021)

Fase 2 - Transformação dos dados: partindo do arquivo csv gerado anteriormente, foi nesta fase que ocorreu o enriquecimento e o processo de triplificação dos dados. A implementação de acordo com o modelo e a ontologia selecionada foi realizada através dos componentes do *plugin* ETL4LOD+. Os componentes *Data Property Mapping* e *Object Property Mapping* foram responsáveis pelo mapeamento das propriedades e valores dos dados provenientes da fase 1 (Item B em amarelo da figura 17). O *NTriple Generator* foi utilizado para a triplificação, gerando ao final do fluxo um arquivo xml com as triplas RDF (Item C em roxo da figura 17). Outro aspecto importante da fase 2 foi a geração da URI do sujeito para

identificação e acesso único ao conjunto de dados (Item A em azul da figura 17). A figura 17 ilustra a modelagem e a saída serializada em um arquivo xml triplificado:

FIGURA 17 – ILUSTRAÇÃO DA TRIPLIFICAÇÃO DOS DADOS NA FASE 2

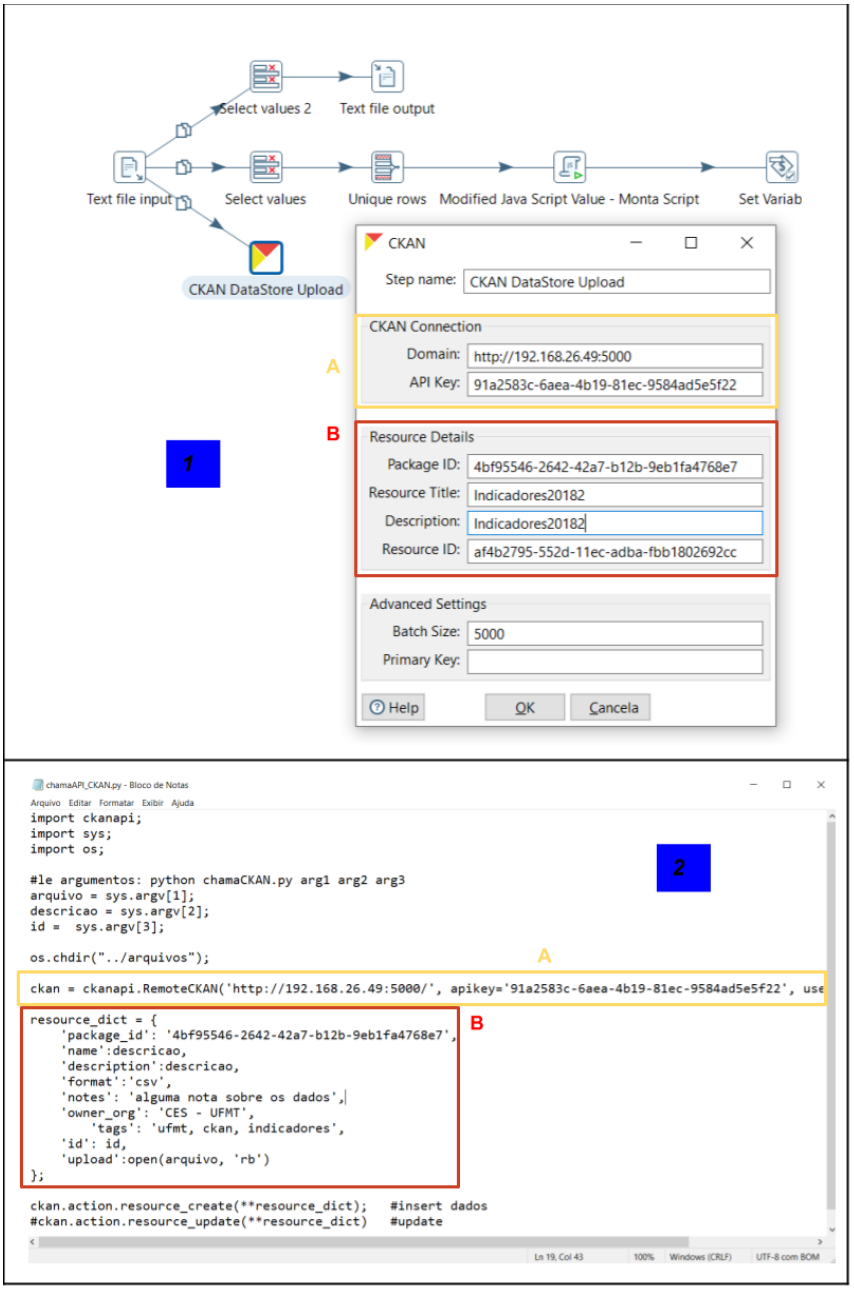


Fonte: Elaborado pelo autor (2021)

Fase 3 - Carga dos dados abertos educacionais: a última fase corresponde a carga dos DAE no catálogo da UFMT. Por meio dos recursos da API do CKAN foi possível a gestão e a publicação do conjunto de dados gerados durante o processo. A URL criada pela ferramenta CKAN foi utilizada para representar o sujeito da tripla RDF e para acesso e identificação única dos dados abertos publicados ao final da fase. Outra característica importante do CKAN foi a possibilidade de controle sobre os metadados através da API. Na chamada do método de carga dos recursos foi possível passar os metadados via parâmetros, agregando descrições ao conjunto de dados publicados.

A figura 18 destaca no item A em amarelo as configurações de acesso da api e em vermelho, no item B, a elaboração dos parâmetros com os metadados nas duas opções de componentes de carga testadas (1 - *plugin* CKAN DataStore Upload e 2 - script em python):

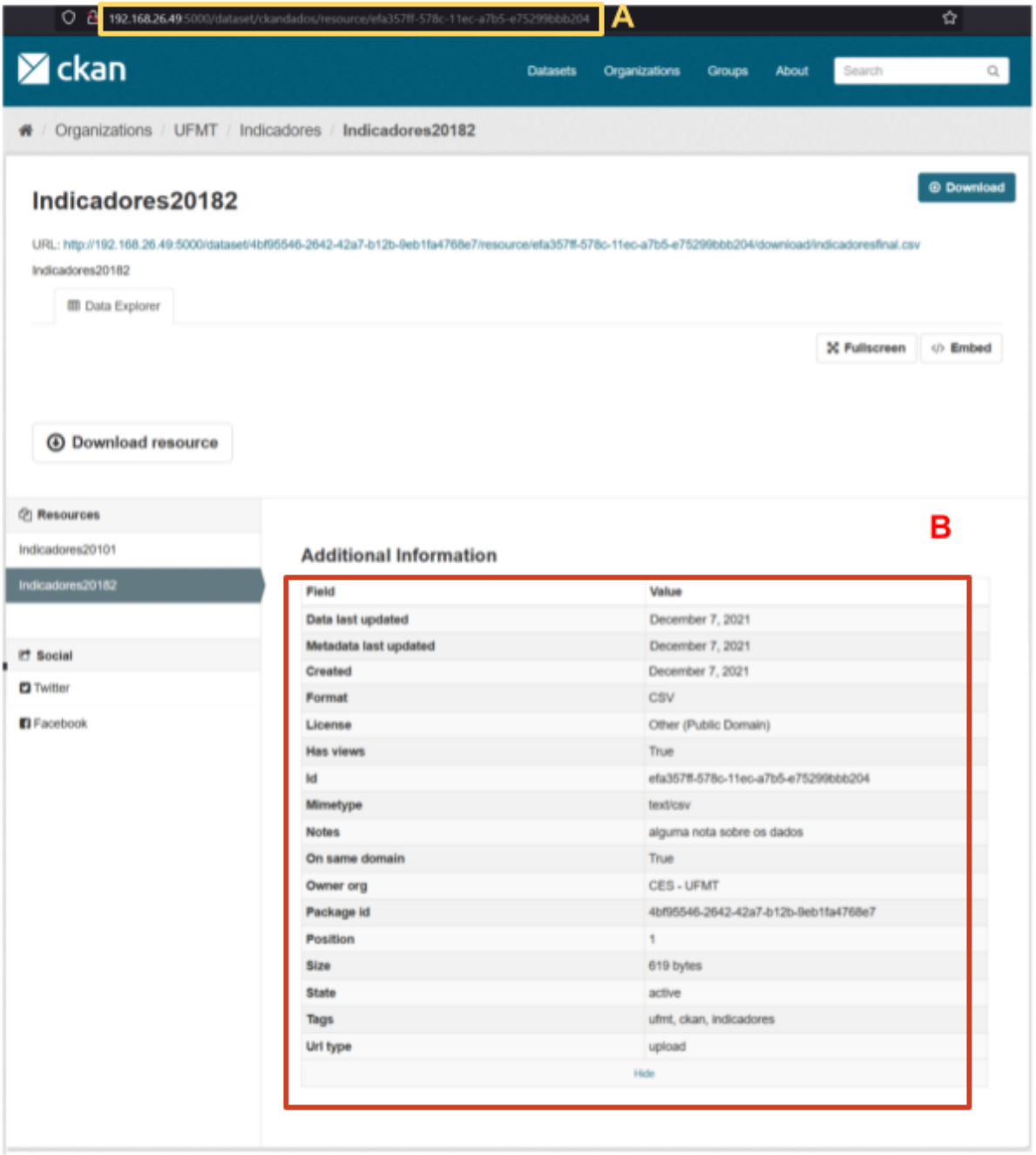
FIGURA 18 – IMPLEMENTAÇÃO DOS COMPONENTES DE CARGA



Fonte: Elaborado pelo autor (2021)

Ao final da etapa 3.4 os dados abertos estão publicados e disponíveis para consulta pública no catálogo, somados dos metadados descritivos e acessíveis através de uma URL única de identificação. A figura 19 apresenta a página final com a URL destacada em amarelo (item A) e os metadados em vermelho (item B):

FIGURA 19 – DAE PUBLICADOS

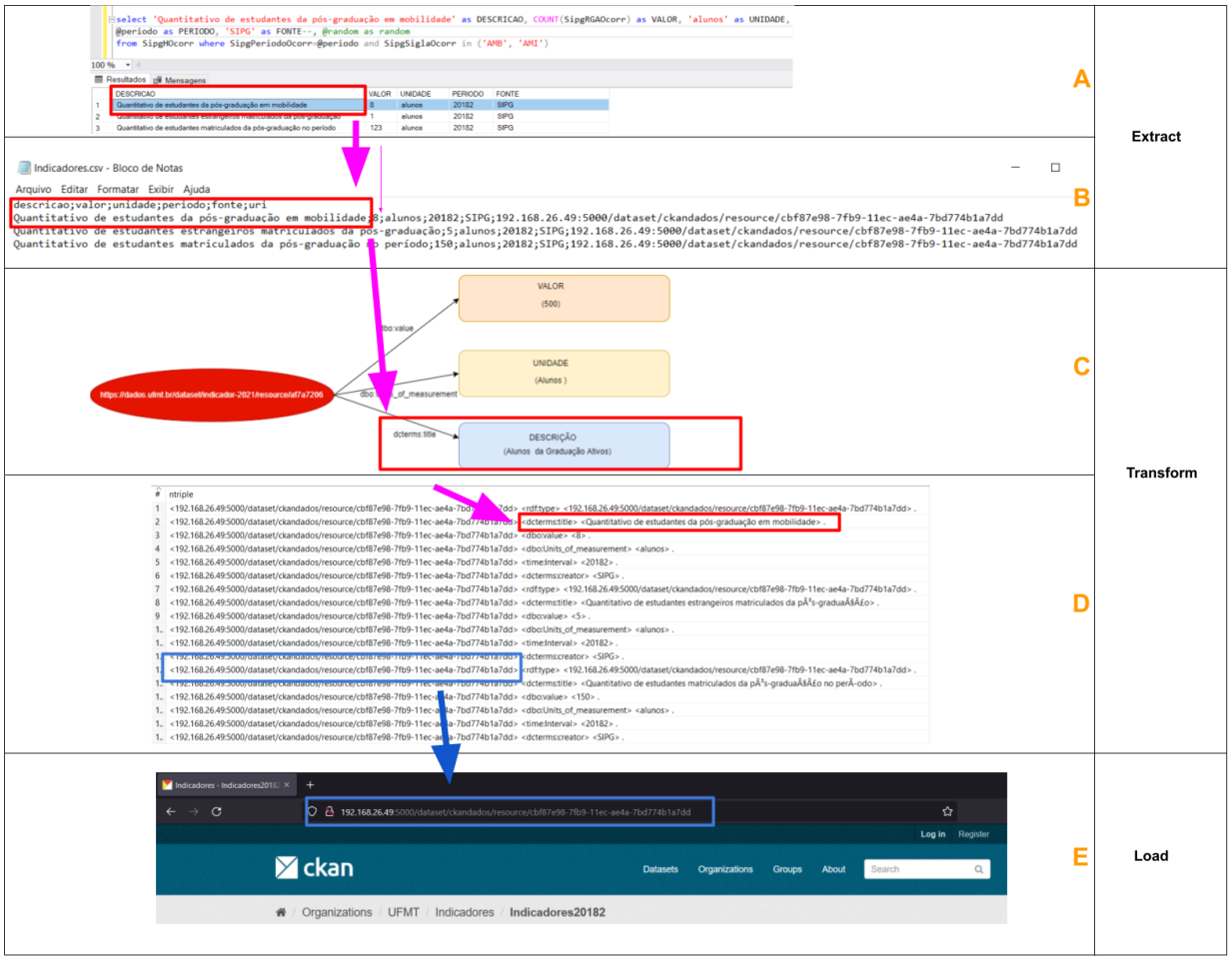


Fonte: Elaborado pelo autor (2021)

Após percorrer as etapas do método proposto para a implementação da aplicação foi possível validar o método com uma demonstração real, com a produção e publicação de dados abertos no contexto da UFMT. Os DAE produzidos pela aplicação partiram de um conjunto

de dados reais extraídos da base de dados relacional da universidade, passando pelo enriquecimento semântico e a triplificação. A figura 20 representa o caminho percorrido pelos dados ao longo de todo o processo, iniciando com a extração por meio da consulta em SQL (item A) e o arquivo em formato csv com os dados públicos brutos (item B). Após a extração, temos a modelagem dos dados e o mapeamento para seus respectivos termos. O item C representa o modelo empregado para a representação dos indicadores educacionais e o item D o resultado da triplicação ao final da fase 2 (RDF serializado em arquivos xml ou json). Ao final da figura 20 temos os dados publicados no portal de catalogação da instituição, com a URL de acesso aos dados indicado pelo retângulo azul (item E):

FIGURA 20 – TRANSFORMAÇÃO DOS DADOS DURANTE PROCESSO



Fonte: Elaborado pelo autor (2021)

Analisando o grau de maturidade e as características dos DAE gerados e disponibilizados no CKAN pela aplicação, constatou-se os seguintes atributos (LÓSCIO et al., 2018): dados liberados na *web* com licença aberta, em formato estruturado, legível por máquina, aberto e com identificadores URI. De acordo com as características apresentadas e da classificação das cinco estrelas de Tim Berners-Lee (BERNERS-LEE, 2006), os dados abertos decorrentes da solução proposta são classificados como dados 4 estrelas, faltando apenas a ligação com outros dados para alcançar a quinta estrela (dados abertos conectados). Desta forma, considerou-se satisfatório o resultado alcançado ao final da solução.

6 CONSIDERAÇÕES FINAIS

As TICs têm o potencial de inovar os serviços públicos, tornando-os mais eficientes, estreitando a relação com os cidadãos através da tecnologia. No Brasil existem leis garantindo o acesso aos dados públicos pela sociedade. Dessa forma, faz-se possível a criação de mecanismos tecnológicos com o objetivo de disponibilizar na *web* os dados abertos governamentais.

A presente dissertação teve por objetivo propor um método para produzir e disponibilizar de forma automatizada dados abertos educacionais na UFMT, visando o desenvolvimento de uma solução inovadora para suprir a lacuna da falta de um processo automatizado para a produção de DAE aplicável na UFMT.

Os objetivos do trabalho foram alcançados com a proposta de um método para captura e compartilhamento de dados abertos e sua comprovação através da solução em ETL que produziu dados 4 estrelas ao final do processo, conforme Tim Berners-Lee (BERNERS-LEE, 2006). Após o desenvolvimento da solução, o método foi verificado via uma oficina com um grupo focal e os resultados da oficina verificaram que a solução proposta atende ao seu objetivo, obtendo um bom nível de aceitação.

Como principais contribuições desta dissertação no campo do mestrado profissional, temos os seguintes resultados:

- Proposição do método original para captura e compartilhamento de dados abertos utilizando uma abordagem ETL, disponibilizando um processo replicável e escalável. O método é composto por uma representação gráfica e uma tabela com a explicação de cada fase do método e as descrições sobre as principais atividades.

- Desenvolvimento da aplicação *desktop* para instanciar o método e o depósito do código-fonte em um repositório reconhecido. Com a utilização de ferramentas de código aberto como o CKAN, Kettle e o *plugin* ETL4LOD foi possível atender há uma demanda real da instituição com a implementação da solução.

- Realização da demonstração do método com um exemplo real comprovando a viabilidade prática do método, com a utilização de um conjunto de dados verdadeiros para a produção e publicação de dados abertos no contexto da UFMT.
- Submissão de um artigo científico em revista com Qualis CAPES adequado.

Durante o desenvolvimento da dissertação algumas limitações foram observadas. A primeira delas foi a utilização apenas do modelo de dados para representar indicadores educacionais. Na educação superior existe uma infinidade de dados abertos passíveis de publicação, a solução proposta limitou-se apenas aos indicadores educacionais. Outra característica limitante do método é o público-alvo. O projeto é voltado para uma equipe de desenvolvimento, gerando dependência de analistas de sistemas com conhecimento nas ferramentas de ETL e de toda a infraestrutura para instalação e configuração do ambiente necessário para rodar o Kettle e o CKAN. A aplicação foi experimentada com um conjunto de dados pequenos, futuramente, a solução deve ser testada com outros volumes de dados oriundos de planilhas ou serviços *web*.

Como trabalho futuro fica o desafio de potencializar o alcance dos dados abertos adicionando conexões com outros dados, criando dados abertos conectados ao final do processo, elevando para 5 estrelas os DAE disponibilizados. No domínio da infraestrutura tecnológica sugere-se a configuração de um servidor SPARQL que permita a hospedagem e consulta de dados no formato RDF como, por exemplo, os servidores Apache Jena ou Virtuoso. Atualmente não existe um servidor SPARQL para consultas na instituição. É indicado também como trabalho futuro o treinamento de uma equipe de desenvolvimento, deixando-os aptos para o suporte e com conhecimento sobre o método e a aplicação.

O levantamento do estado da arte indicou que ainda existem lacunas nas abordagens e padronizações existentes para publicação de dados abertos governamentais no cenário brasileiro, enfatizando a importância de criar soluções automatizadas para a produção de dados abertos em formatos apropriados, passíveis de serem localizados por agentes de *software* ou pessoas, contribuindo para a transparência e eficiência do setor público.

Cabe salientar que o método proposto é inovador frente à alimentação manual do catálogo de dados abertos da UFMT, podendo gerar um ganho de eficiência e transparência em relação às publicações pontuais realizadas pelas unidades da instituição, apresentando como característica principal a automatização do processo, de forma escalável e replicável.

O método não é um modelo engessado, mas sim uma sequência de etapas orientando a instrução a ser seguida, que poderá ser aplicada de acordo com as particularidades encontradas. A implementação do método na ferramenta ETL selecionada é genérico, extensível, flexível e escalável, possibilitando o uso de tecnologias livres como o processo ETL e ferramentas de código aberto. Cada etapa do método é executável pelas diversas ferramentas disponíveis na atualidade, como o Talend, SQL Server Integration Services

(SSIS), Pentaho Data Integration (Kettle), entre outros, adaptando-se às especificidades de cada demanda, conjunto de dados ou infraestrutura tecnológica (servidores de banco de dados, webs etc).

Por fim, cabe ressaltar que acredita-se ter contribuído de forma inovadora para a gestão de Dados Abertos Educacionais, tanto para a UFMT quanto para a sociedade em geral, com o registro do método desenvolvido via esta dissertação.

REFERÊNCIAS

ALCANTARA, W.; BANDEIRA, J.; BARBOSA, A.; LIMA, A.; ÁVILA, T.; BITTENCOURT, I.; ISOTANI, S. Desafios no uso de dados abertos conectados na educação brasileira. *In*: WORKSHOP DE DESAFIOS DA COMPUTAÇÃO APLICADA À EDUCAÇÃO (DESAFIE!), 4., 2015, Recife. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2015. p. 11-20. DOI: https://doi.org/10.5753/desafie.2015.10036. Disponível em: https://sol.sbc.org.br/index.php/desafie/article/view/10036. Acesso em: 15 jan. 2021.

ALENCAR, A.; XAVIER, D.; CHAVES, L.; SOUZA, D. Publicação e consumo de dados abertos conectados acadêmicos. *Revista Principia - Divulgação Científica e Tecnológica do IFPB*, [*S.l.*], n. 42, p. 136-145, ago. 2018. ISSN 2447-9187. Disponível em: https://periodicos.ifpb.edu.br/index.php/principia/article/view/1988. Acesso em: 8 dez. 2020.

ALENCAR, A.; XAVIER, D.; CHAVES, L. C.; SOUZA, D. Publicação e consumo de dados abertos conectados acadêmicos. *In*: *SIMPÓSIO BRASILEIRO DE BANCO DE DADOS - SBBD*, 32., 2017. Disponível em: https://periodicos.ifpb.edu.br/index.php/principia/article/view/1988. Acesso em: 15 jan. 2021.

ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciênc. saúde coletiva*, Rio de Janeiro, v. 16, n. 7, p. 3061-3068, jul. 2011. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232011000800006&lng=pt&nrm=iso>. Acesso em: 8 dez. 2020.

ARAÚJO L. R.; SOUZA J. F. Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. *Revista Eletrônica de Sistemas de Informação*, [*S.l.*], v. 10, 2011. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/880. Acesso em: 15 dez. 2020.

ATTARD, J. *et al*. A systematic review of open government data initiatives. *Government Information Quarterly*, v. 32, n. 4, p. 399-418, 2015.

BANDEIRA, J.; ÁVILA, T; ALCANTARA, W.; BARBOSA, A.; BITTENCOURT, I.; ISOTANI, S. Dados abertos conectados para a Educação. *Jornada de Atualização em Informática na Educação*, [*S.l.*], p. 47-69, out. 2015. ISSN 23167734. Disponível em: https://www.br-ie.org/pub/index.php/pie/article/view/3551. Acesso em: 10 jan. 2021.

BERBERIAN, C.; MELLO, P.; CAMARGO, R. Governo Aberto: A tecnologia contribuindo para maior aproximação entre o Estado e a Sociedade. *Revista TCU*, [*S.l.*], n.131, p.30-39, set./dez. 2014. Disponível em: https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/60. Acesso em: 7 dez. 2020.

BERNERS-LEE, T. Linked data. *World Wide Web Consortium (W3C)*, 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 9 abr. 2021.

BRANDT, M. B.; VIDOTTI, S. A. B. G.; SEGUNDO, J. E. S. Modelo de dados abertos conectados para informação legislativa. *Informação & Sociedade: Estudos*, v. 28, n.2, 2018. Disponível em: https://periodicos.ufpb.br/index.php/ies/article/view/37979. Acesso em: 15 jan. 2021.

BRASIL. Lei nº 12.527 de 18 de novembro 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 18 nov. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 10 dez. 2020.

CARVANO, L. M. F. *A utilização de dados públicos abertos na construção de um Data Warehouse:* A construção de um repositório estatísticas educacionais públicas brasileiras. 2018. Dissertação (Mestrado em Gestão de Informação) - NOVA Instituto Superior de Estatística e Gestão de Informação Universidade Nova, Lisboa, 2018.

COELHO, T. R.; SILVA, T. A. B.; CUNHA, M. A.; TEIXEIRA, M. A. C. Transparência governamental nos estados e grandes municípios brasileiros: uma “dança dos sete véus” incompleta? *Cadernos Gestão Pública e Cidadania*, v.23, n.75, 2018. Disponível em: http://bibliotecadigital.fgv.br/ojs/index.php/cgpc/article/view/73447. Acesso em: 11 dez. 2020.

DADOS.GOV.BR. Portal Brasileiro de Dados Abertos. Governo Federal. *O que são dados abertos?* Disponível em: https://dados.gov.br/pagina/dados-abertos. Acesso em: 25 de mar. 2021.

DADOS.GOV.BR. Portal Brasileiro de Dados Abertos. Governo Federal. *Wiki da INDA*. Página modificada em 19 ago. 2021, 11:57 por Administrator. Disponível em: https://wiki.dados.gov.br/. Acesso em: 25 mar. 2021.

DALMORO, M.; VIEIRA, K. M. DILEMAS NA CONSTRUÇÃO DE ESCALAS TIPO LIKERT: O NÚMERO DE ITENS E A DISPOSIÇÃO INFLUENCIAM NOS RESULTADOS? 3. ed. [S.l.]: *REVISTA GESTÃO ORGANIZACIONAL*, 2013. 36-40 p. Disponível em: https://bell.unochapeco.edu.br/revistas/index.php/rgo/article/view/1386/1184. Acesso em: 11 dez. 2021.

DIAGRAMS.NET. *Security-first diagramming for teams*. [c2005-2021]. Disponível em: diagrams.net. Acesso em: 11 jul. 2021.

DUTRA, C. C.; LOPES, K. M. G. Dados abertos: uma forma inovadora de transparência. *In*: CONGRESSO CONSAD DE GESTÃO PÚBLICA, 6., 2013, Brasília-DF. *Anais* [...]. Brasília-DF: Centro de Convenções Ulysses Guimarães, 2013. Disponível em: http://www.sgc.goias.gov.br/upload/arquivos/2014-09/dados-abertos---uma-forma-inovadora-de-transparEncia.pdf. Acesso em: 8 fev. 2021.

EPING. *Padrões de Interoperabilidade de Governo Eletrônico - ePING*. Disponível em: http://eping.governoeletronico.gov.br/. Acesso em: 25 de mar. 2021.

FERREIRA, M. C. O direito fundamental à informação pública como condição de efetividade da democracia no Estado Democrático de Direito e o segredo como restrição. *Jus Navigandi*, 2017. Disponível em: https://jus.com.br/artigos/58442/o-direito-fundamental-a-informacao-publica-como-condicao-de-efetividade-da-democracia-no-estado-democratico-de-direito-e-o-segredo-como-restricao. Acesso em: 7 jan. 2021.

GONÇALVES, B. A.; GAMA, K. S. Transparência e dados abertos do recife: uma estratégia bem-sucedida de publicação. *In*: CONFERÊNCIA LUSO-BRASILEIRA DE ACESSO ABERTO, 9., 2018. Disponível em: https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1901. Acesso em: 8 dez. 2020.

GOV.BR. Ministério da Economia. Governo Digital. *Dados Abertos Governamentais*. Disponível em: https://www.gov.br/governodigital/pt-br/dados-abertos/dados-abertos-governamentais. Acesso em: 25 de mar. 2021.

ISOTANI, S.; BITTENCOURT, I. I. *Dados abertos conectados*. São Paulo: Novatec, 2015.

JAMIL, G. L.; NEVES, J. T. R. A era da informação: considerações sobre o desenvolvimento das tecnologias da informação. *Revista Perspectivas em Ciência da Informação*, v.5, 2000. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/23309/18844. Acesso em: 10 jul. 2021.

KLEIN, R. H.; KLEIN, D. C. B.; LUCIANO, E. M. Identificação de Mecanismos para a Ampliação da Transparência em Portais de Dados Abertos: Uma Análise no Contexto Brasileiro. *Cadernos EBAPE.BR*, v.16, n. 4, 2018. Disponível em: http://bibliotecadigital.fgv.br/ojs/index.php/cadernosebape/article/view/73241. Acesso em: 12 jan. 2021.

LÓSCIO, B. F.; BURLE, C.; OLIVEIRA, M. I. S.; CALEGARI, N. Fundamentos para publicação de dados na Web. *Centro de Estudos sobre Tecnologias Web (Ceweb.br)*, 2018.

Disponível em: https://ceweb.br/publicacao/livro-fundamentos-dados-web/. Acesso em: 13 mar. 2020.

MACHADO, S. M.; FIDELIS, A. C. F.; CARRARO, I. R.; OLEA, P. M. Pesquisa Científica: Conhecimento e Percepção dos Acadêmicos de Administração em Caxias do Sul. *Revista E-Tech: Tecnologias Para Competitividade Industrial*, 2016. Disponível em: https://etech.sc.senai.br/edicao01/article/view/787. Acesso em: 14 mai. 2021.

MACIEL, C. *Um método para mensurar o grau de maturidade na tomada de decisão e-democrática*. 2008. Tese (Doutorado em Computação) - Universidade Federal Fluminense, Niterói, 2008.

MARTINS, L. C. B. *Proposta de arquitetura de publicação automatizada de dados abertos conectados utilizando meta-dados e ontologias*. 2018. Dissertação (Mestrado Profissional em Computação Aplicada) - Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2018.

MARTINS, L. C. B.; VICTORINO, M. C.; HOLANDA, M.; GHINEA, G; GRØNLI, T. UnBGOLD: UnB government open linked data - Semantic enrichment of open data tool. *In:* INTERNATIONAL CONFERENCE ON MANAGEMENT OF DIGITAL ECOSYSTEMS, 10., 2018. *Proceedings* […], 2018. Disponível em: https://bura.brunel.ac.uk/bitstream/2438/18443/1/FullText.pdf. Acesso em: 8 mai. 2021.

MENDONÇA, P. G. A.; MACIEL, C.; VITERBO, J. Visualizing aedes aegypti infestation in urban areas: A case study on open government data mashups. *Information Polity*, v. 20, n. 2, 3, p. 119-134, 2015.

MIZOGUCHI, R. Tutorial on ontological engineering: part 3: Advanced course of ontological engineering. *New Generation Computing,* v. 22, n. 2, p. 198-220, 2004.

MORAES NETO, A. J.; SILVA, C. E.; ANJOS, W. F.; DORÇA, F. A. Uma abordagem baseada em dados abertos conectados e chatbot para disponibilizar o catálogo de cursos da rede federal de educação profissional, científica e tecnológica. *In*: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - CBIE , 9., 2020. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/12882/12736. Acesso em: 8 jan. 2021.

OLIVEIRA, E. C.; GUIMARÃES, J. V. M.; COSTA, S. S. Migrando dos dados abertos para dados conectados: uma proposta para a Universidade Federal do Maranhão. *In:* JORNADA DE INFORMÁTICA DO MARANHÃO - JIM, 7, 2018. Disponível em: http://sistemas.deinf.ufma.br/anaisjim/artigos/2018/201818.pdf. Acesso em: 8 junho 2021.

OPEN KNOWLEDGE BRASIL. *Por que open*. 2020. Disponível em: https://www.ok.org.br/dados-abertos/. Acesso em: 11 dez. 2020.

PENTAHO. *Data Integration - Kettle*, 2021. Disponível em: https://community.hitachivantara.com/s/article/data-integration-kettle. Acesso em: 11 de fev. 2021.

PENTEADO, B. E. *Modelo de infraestrutura para publicação de dados abertos governamentais conectados de qualidade*. 2020. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

PENTEADO, B. E.; BITTENCOURT, I. I.; ISOTANI, S. Análise exploratória sobre a abertura de dados educacionais no Brasil: como melhorar o ecossistema de dados na Web? *Revista Brasileira de Informática na Educação – RBIE*, [*S.l.*], v. 27, n.1, p.175-195, 2019a. DOI:10.5753/RBIE.2019.27.01.175. Disponível em: https://repositorio.usp.br/directbitstream/e8a28063-1622-445c-9c7a-11545b03e24b/2937558.pdf. Acesso em: 8 jan. 2021.

PENTEADO, B. E.; BITTENCOURT, I. I.; ISOTANI, S. Metaprocesso para transformação de dados educacionais em dados conectados. *In*: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - CBIE, 8., 2019. *Anais*[...], 2019b. Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/8893. Acesso em: 10 jan. 2021.

PENTEADO, B. E.; BITTENCOURT, I. I.; ISOTANI, S. Modelo de referência para dados abertos educacionais em nível macro. *In*: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - CBIE, 8., 2019. *Anais*[...], 2019c. Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/8914. Acesso em: 10 jan. 2021.

PENTEADO, B. E.; MALDONADO, J. C.; ISOTANI, S. Process model with quality control for the production of high quality linked open government data. *IEEE LATIN AMERICA TRANSACTIONS*, v. 19, n. 3, 2021. Disponível em: https://latamt.ieeer9.org/index.php/transactions/article/view/3501. Acesso em: 10 jan. 2021.

PRODANOV, C. C.; FREITAS, E. C. *Metodologia do trabalho Científico*: métodos e técnicas da pesquisa e do Trabalho Acadêmico. Novo Hamburgo: Feevale, 2013.

RAUTENBERG, S.; MOTYL, S. K.; BURDA, A. C.; SILVÉRIO, A.; MOURA, F. M. Dados abertos conectados e gestão do conhecimento: estudos de caso cientométricos em uma universidade brasileira. *Perspectivas em Ciência da Informação*, v. 22, n. 3, 2017. Disponível em: https://www.scielo.br/j/pci/a/KykXkxTPkz369RZjZCfmjnR/?format=html&lang=pt. Acesso em: 18 jan. 2021.

REIS JR., C.; MARTINS, L.; VICTORINO, M.; HOLANDA, M. Modelo de dados de proveniência para uma arquitetura de dados abertos governamentais. *In*: WORKSHOP DE TRANSPARÊNCIA EM SISTEMAS, 7., 2019. *Anais*[...], 2019. Disponível em: https://sol.sbc.org.br/index.php/wtrans/article/view/6437/6333. Acesso em: 8 jan. 2021.

SANTOS, O. A. R. *Minha escola transparente*: uma análise comparativa do uso de dados governamentais abertos na educação básica no Brasil e Inglaterra. 2014. Dissertação (Mestrado Profissional em Administração Pública) - EBAP, FGV, Rio de Janeiro, 2014.

SANTOS, S. S. *Um processo para conversão e publicação de dados para modelo RDF seguindo os princípios de linked data*. 2016. Trabalho de Conclusão de Curso (Bacharel em Sistemas de Informação) - Universidade Federal do Ceará, Ceará, 2016.

SILVA, J. F. C. *ETL4LOD+*: evolução do suporte ao ciclo de publicação de dados conectados. 2018. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação) - Departamento de Ciência da Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018.

SILVA, M. L.; COSTA, S. S.; LIMA, T. S.; SILVA, T. A. Um estudo exploratório sobre dados abertos em universidades. *In:* JORNADA DE INFORMÁTICA DO MARANHÃO - JIM, 7., 2016. *Anais*[...], 2016. Disponível em: https://www.researchgate.net/profile/Sergio-Souza-Costa/publication/310233815_Um_estudo_exploratorio_sobre_dados_abertos_em_Universidades/links/582a4d8508ae004f74ae537a/Um-estudo-exploratorio-sobre-dados-abertos-em-Universidades.pdf. Acesso em: 8 junho 2021.

SILVEIRA, R. N. *Método para rotular ligações semânticas na web de dados*. 2021. Dissertação (Mestrado em Ciências em Sistemas e Computação) - Programa de Pós-graduação em Sistemas e Computação, Instituto Militar de Engenharia, Rio de Janeiro, 2021.

TCU. Tribunal de Contas da União. Portal do TCU. *Cinco motivos para a abertura de dados na administração pública*, 2015. Disponível em: https://portal.tcu.gov.br/biblioteca-digital/cinco-motivos-para-a-abertura-de-dados-na-administracao-publica.htm. Acesso em: 25 mar. 2021.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. *Ciência da Informação*, v. 48, n. 3, 2019. Disponível em: http://repositorio.utfpr.edu.br/jspui/bitstream/1/4812/1/dadosabertosdesafiosdisponibilizacaoweb.pdf. Acesso em: 11 jan. 2021.

VICTORINO, M. C.; MARTINS, L.; HOLANDA, M.; FONSECA, R. Arquitetura de publicação de dados abertos conectados governamentais da universidade de Brasília. *Encontros Bibli*: Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 25, p. 1-25. 2020. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e67665. Acesso em: 8 mai. 2021.

UFMT. Plano de Dados Abertos. Universidade Federal de Mato Grosso - PDA. *Transparência*. Disponível em: https://www.ufmt.br/pagina/transparencia/947. Acesso em: 22 de mar. 2021.

APÊNDICE A

Termo de Consentimento Livre e Esclarecido

Você está sendo convidado(a) para participar, como voluntário(a), da pesquisa de mestrado intitulada "Proposta de um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT", desenvolvida por Fabio Antonio Rodrigues, sob orientação do Professor Dr. Cristiano Maciel do Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação - PROFNIT.

O estudo tem por objetivo geral propor um método para coleta e transformação de dados brutos em Dados Abertos Educacionais, visando a publicação em um catálogo de dados abertos e o desenvolvimento de uma aplicação desktop em ETL para instanciar o método proposto no contexto da UFMT.

A sua participação refere-se a etapa de verificação do método e da aplicação. A qualquer momento você poderá desistir de participar, sem penalização alguma ou prejuízo e não terá nenhum custo ou bonificação (vantagem financeira).

Haverá total sigilo e anonimato quanto aos dados e informações prestados.

Em caso de dúvida, você poderá entrar em contato com o pesquisador através do e-mail: fabio.rodrigues@ufmt.br .

Esta pesquisa tem a duração média de 1 hora e 30 minutos e se dividirá em três etapas que serão apresentados resumidamente a seguir:

- Assistir a apresentação do método e da aplicação;

- Participar de um momento de conversa e esclarecimento das dúvidas;

- Responder a um questionário criado pela ferramenta do Google Forms, o qual consiste em perguntas relacionadas ao entendimento sobre dados abertos, método e a aplicação.

APÊNDICE B

Proposta de um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

Questões sobre o perfil do participante

Data de hoje (data do preenchimento): *

Month, day, year

Idade (em anos): *

Short answer text

Titulação Máxima Completa: *

- Graduado(a)
- Especialista
- Mestre(a)
- Doutor(a)

Área de atuação profissional: *

- Público
- Privado
- Público e Privado

Qual o seu conhecimento sobre Dados Abertos?

Long answer text

APÊNDICE C

Proposta de um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

Questões específicas sobre o sentimento do participante

A solução está de acordo com o PDA da UFMT no quesito de propor uma forma automatizada para publicação dos dados abertos, com ganhos de eficiência em comparação a extrações pontuais. *

- Concordo Totalmente
- Concordo Parcialmente
- Não Concordo Nem Discordo
- Discordo Parcialmente
- Discordo Totalmente

O método/aplicação são escaláveis para outros conjuntos de dados ou instituições. *

- Concordo Totalmente
- Concordo Parcialmente
- Não Concordo Nem Discordo
- Discordo Parcialmente
- Discordo Totalmente

O método com foco na abordagem ETL e a ferramenta open source Kettle atendem os requisitos tecnológicos para geração de Dados Abertos Governamentais. *

- Concordo Totalmente
- Concordo Parcialmente
- Não Concordo Nem Discordo
- Discordo Parcialmente
- Discordo Totalmente

A aplicação e o método são de simples compreensão. *

- Concordo Totalmente
- Concordo Parcialmente
- Não Concordo Nem Discordo
- Discordo Parcialmente
- Discordo Totalmente

A solução atendeu ao objetivo de propor um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT. *

- Concordo Totalmente
- Concordo Parcialmente
- Não Concordo Nem Discordo
- Discordo Parcialmente
- Discordo Totalmente

Justifique a sua resposta acima, comente: *

Long answer text

APÊNDICE D

Proposta de um método para captura e compartilhamento de dados abertos no contexto dos sistemas da UFMT

Avaliação do grupo focal

Como você avalia a experiência do grupo focal? *

- Muito boa
- Boa
- Neutra
- Ruim
- Péssima

Deixe a sua opinião ou comentário sobre a dinâmica do grupo focal:

Long answer text