Um método para captura e compartilhamento de dados abertos educacionais via um processo ETL

Fabio A. Rodrigues[1,3], Cristiano Maciel[2,3]

[1]Secretaria de Tecnologia da Informação - Universidade Federal de Mato Grosso
UFMT - Cuiabá – MT – Brasil

[2] Instituto de Computação - Universidade Federal de Mato Grosso
UFMT - Cuiabá – MT – Brasil

[3]Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a inovação, PROFNIT – Ponto Focal Cuiabá

{fabio.rodrigues,cristiano.maciel}@ufmt.br

Abstract. *This article discusses a method for capturing and sharing open educational data in an educational institution via an Data Extraction, Transformation and Loading process, aimed at publishing quality open data in a format suitable for consumption by both software agents and humans. In addition to the proposed method, an application was implemented using the open source tool Kettle, validating the method with a real dataset.*

Resumo. *Este artigo discute sobre um método para captura e compartilhamento de dados abertos educacionais em uma instituição de ensino, a UFMT, via um processo de Extração, Transformação e Carga de dados, visando a publicação de dados abertos de qualidade em um formato apropriado para consumo tanto por agentes de* software *como humanos. Além do método proposto, uma aplicação foi implementada utilizando a ferramenta de código aberto Kettle, validando o método com um conjunto de dados reais.*

1. Introdução

Mais do que a mídia convencional, as TICs (Tecnologia da Informação e Comunicação) e a internet têm contribuído para o acesso mais rápido e integral às informações públicas. 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”.

No Brasil o acesso aos dados públicos 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].

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". Quando disponibilizados gratuitamente por entidades públicas são denominados dados abertos governamentais. No ambiente educacional recebem o nome de dados abertos educacionais – DAE.

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]. De forma complementar, Maciel [2008, p. 5] adverte da "carência de soluções eficientes, efetivamente disponíveis e inovadoras para alavancar a participação popular[...]". Neste sentido, disponibilizar dados abertos educacionais é salutar para as instituições e para a sociedade.

Tomando as evidências dos estudos mencionados, o presente trabalho propõe um processo automatizado, por meio de um método replicável, para prover o acesso a informações educacionais no contexto de uma instituição de ensino superior (IES), a Universidade Federal de Mato Grosso, via um processo de ETL (Extração, Transformação e *Carga de dados, do inglês Extract - Transform - Load*).

2. Bases téoricas

Deseja-se com a fundamentação deste capítulo buscar referências de soluções para geração, representação e publicação de dados abertos governamentais no cenário nacional. Para o alcance do objetivo os 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. A proposta 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 um metaprocesso composto por cinco fases: especificação, modelagem, conversão, publicação e exploração [PENTEADO; BITTENCOURT;I SOTANI, 2019b]. Em trabalho mais recente sobre o tema, foi adicionada uma sexta fase: manutenção [PENTEADO; MALDONADO; ISOTANI, 2021].

Para orquestrar o processo para geração e publicação de dados abertos governamentais temos a abordagem ETL (extrair, transformar e carregar os dados). A abordagem ETL, conduzindo o *workflow* de geração e publicação dos dados abertos conectados 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].

3. Método proposto

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 instituição, transformando-os em dados abertos educacionais de qualidade no formato RDF, de forma a deixá-los prontos para publicação, seguindo um metaprocesso (ou um *workflow*) na abordagem ETL.

O método leva em consideração aspectos como a 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. Na Figura 1 encontra-se a representação do método por fase:

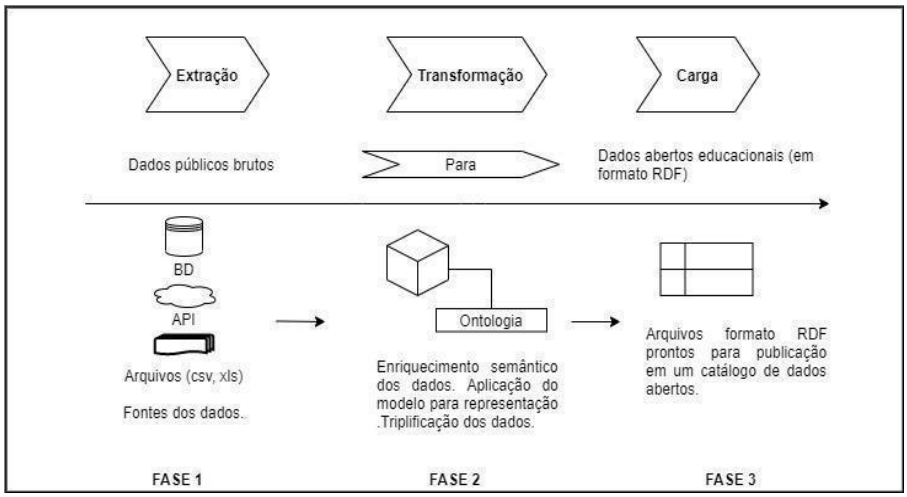

Figura 1. Arquitetura do método

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.

O método proposto é composto por 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 2).

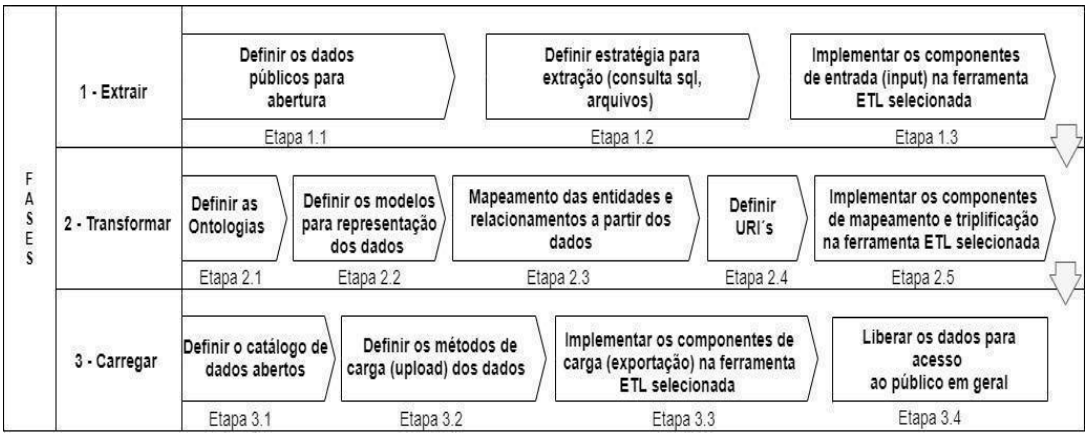

Figura 2. Esquema geral do método

O Quadro 1 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 sequência:

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

Quadro 1. Resumo do método

4. Utilização do método: aplicação *Desktop* em ETL

Esta seção detalha a demonstração do método com um exemplo real, 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.

O centro da solução/aplicação 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).

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 - Definir os dados públicos para abertura: usualmente uma IES possui diversos sistemas cujos dados estão distribuídos em diversas bases de dados, possuindo dados públicos juntos com informações confidenciais. Nesta etapa são selecionados os dados públicos para extração destas diversas fontes possíveis. Vamos utilizar como exemplo indicadores educacionais, tais como o quantitativo de estudantes da pós-graduação em mobilidade, quantitativo de diplomas emitidos por ano ou o quantitativo de estudantes matriculados.

Etapa 1.2 - Definir a estratégia para extração: definição da forma de extrair os dados de acordo com o tipo da fonte de dados (exemplos: banco de dados, planilhas, oriundos de uma api etc). Os conjuntos de dados selecionados neste trabalho estão armazenados em um banco de dados relacional. A estratégia de extração consiste em consultas SQL (SQL *queries*).

Etapa 1.3 - Implementar os componentes de entrada: após identificadas as fontes e os dados selecionados, o próximo passo é implementar os componentes de entrada da ferramenta ETL. A diversidade dos componentes de *input* disponíveis nas ferramentas de ETL possibilitam um leque de opções para a extração dos dados, permitindo minerar dados de arquivos texto, csv, json, diversos banco de dados, entre outros. Na Figura 3, encontram-se o *transformation* de extração e os componentes de *input* (com destaque para o *Table Input* com a consulta SQL).

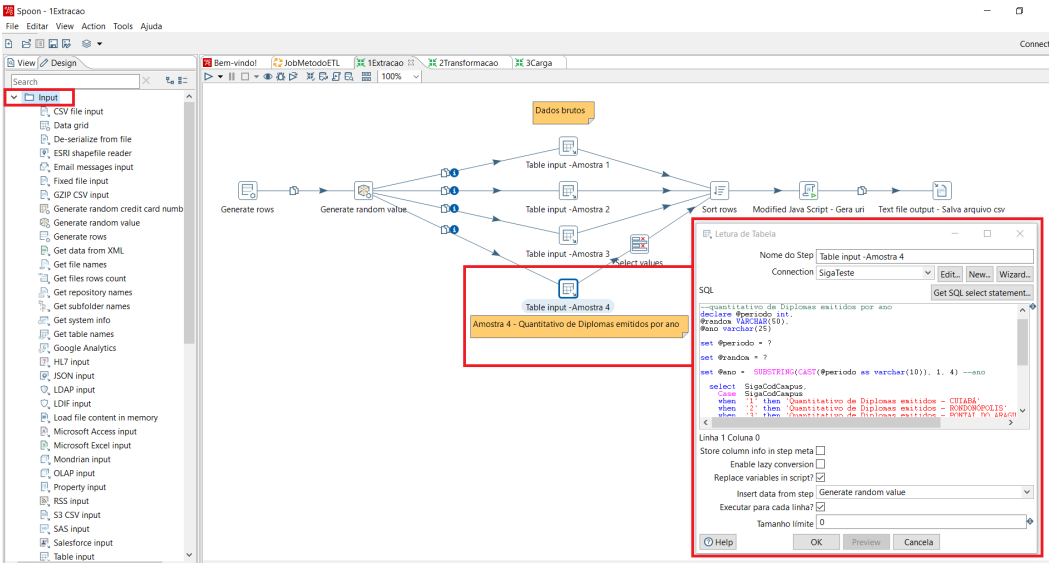

Figura 3. Extração dos dados no Kettle

Etapa 2.1 - Definir as ontologias: na fase 2 inicia-se o processo de enriquecimento e transformação. São selecionadas as ontologias para representação e anotação semântica dos dados originais (Quadro 2).

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/>

Quadro 2. Ontologias

Etapa 2.2 – Definir os modelos para representação dos dados: nesta etapa os dados são modelados conforme os modelos de representação determinados, com o uso do modelo em RDF para exprimir a definição dos dados por meio de um conjunto de triplas, com o sujeito (relação chamada predicado) e o nó objeto (sujeito, predicado, objeto) (ver Figura 4).

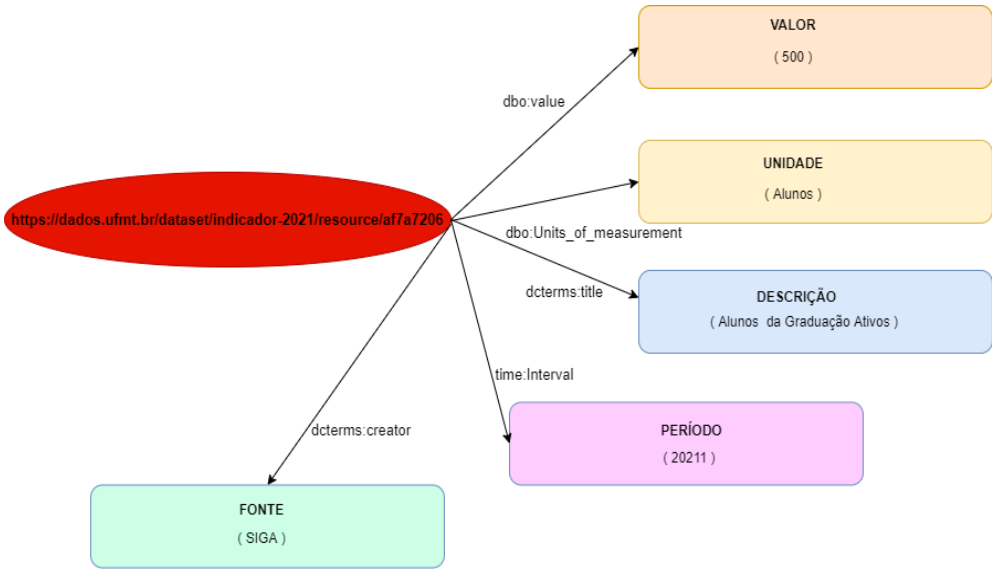

Figura 4. Modelo para representação dos dados

Etapa 2.3 - Mapeamento das entidades: 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 3 apresenta o resultado da etapa.

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

Quadro 3. Mapeamento entre campos e termos

Etapa 2.4 – Definir URI: foi adotado o padrão do catálogo de dados abertos CKAN, com a ferramenta criando uma URL em http única para cada conjunto de dados publicados. Vejamos este exemplo de URL gerada pelo CKAN (Figura 5):

Figura 5. URL gerada pelo CKAN

Etapa 2.5 – Implementar os componentes de transformação: 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 6 apresenta os componentes de transformação.

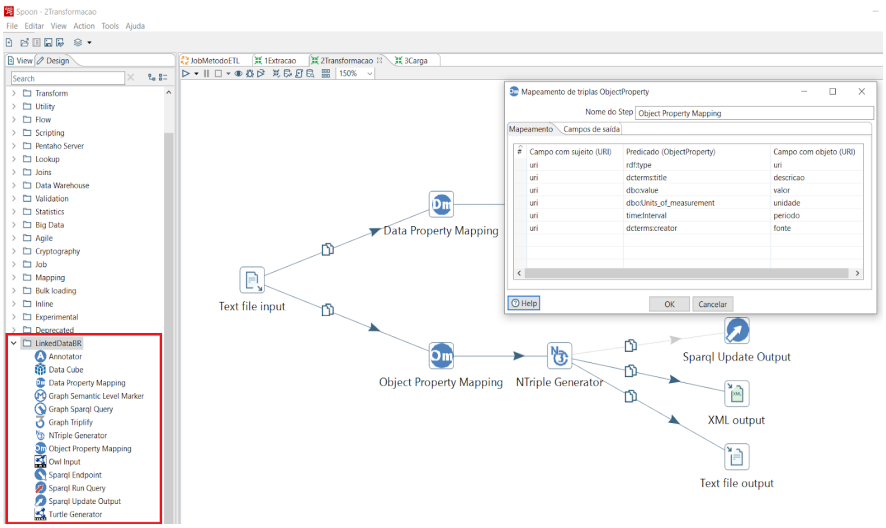

Figura 6. Componentes do ETL4LOD

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 o CKAN foi a ferramenta de catalogação selecionada.

Etapa 3.2 – Definir os métodos de carga: 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. Para o exemplo, a abordagem selecionada foi a chamada à API do CKAN via o componente CKAN *DataStore Upload* ou um *script* em python rodado via Kettle.

Etapa 3.3 – Implementar os componentes de carga: 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 7 apresenta a configuração do componente.

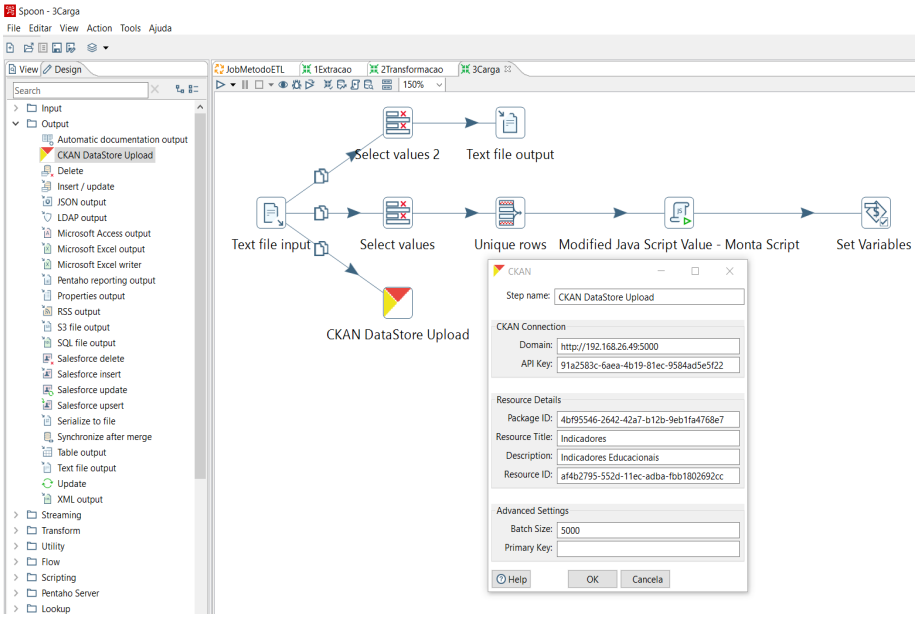

Figura 7. CKAN DataStore Upload

Etapa 3.4 – Liberar os dados para acesso: após a carga no catálogo, os dados abertos estão prontos para consulta pública no portal da instituição, com metadados (Figura 8, item B) e URL de identificação única gerada durante o processo (Figura 8, item A).

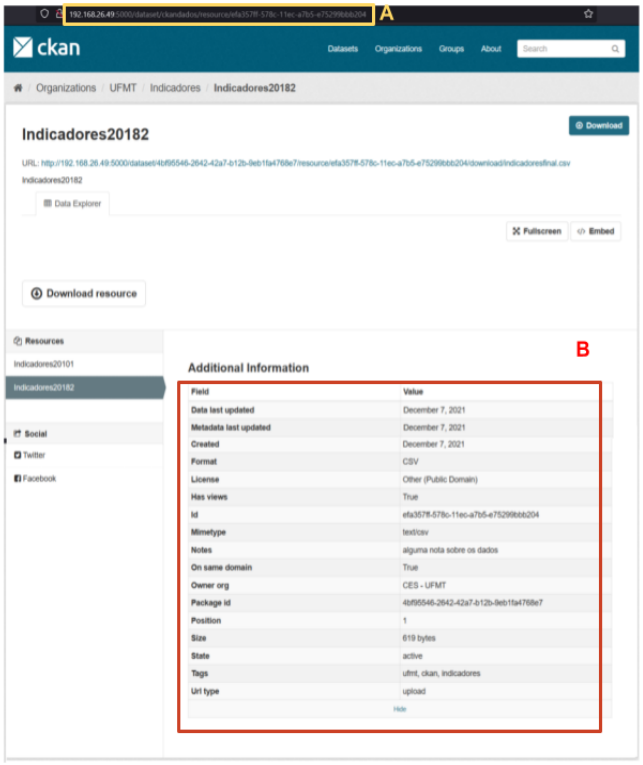

Figura 8. DAE publicados

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.

As Figuras 9 e 10 representam respectivamente o processo de modelagem (do modelo conceitual para a representação em tripla RDF) e a transformação dos dados durante o processo ETL, finalizando com os dados abertos gerados publicados no catálogo de dados abertos acrescidos de metadados e com a URL de identificação única (item E):

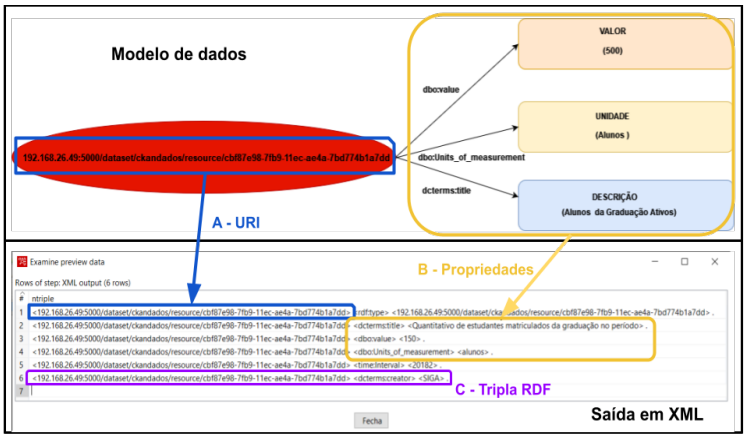

Figura 9. Ilustração da triplificação dos dados na fase 2

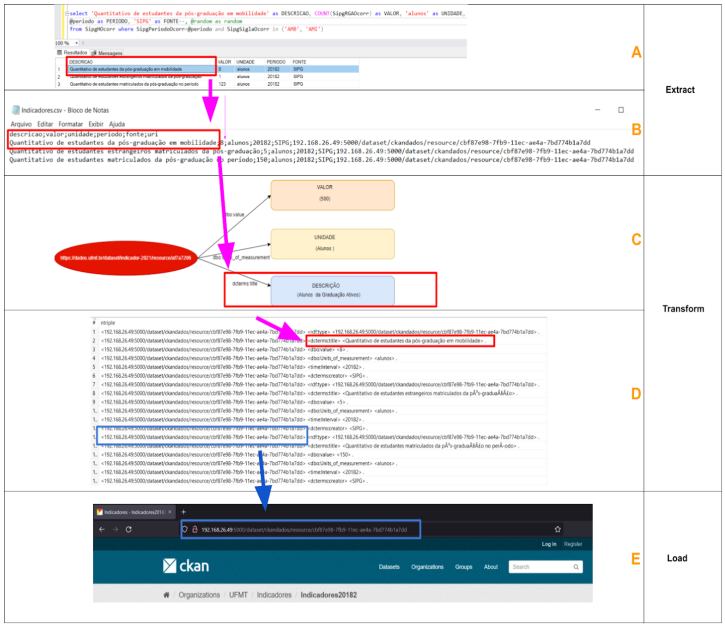

Figura 10. Transformação dos dados durante processo

5. Considerações Finais

O presente trabalho teve por objetivo a proposição de um método para produzir e disponibilizar de forma automatizada dados abertos educacionais, visando o desenvolvimento de uma solução inovadora, apresentando um processo automatizado para a produção de DAE via um processo ETL.

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 (disponível na *web* com licença aberta, em formato estruturado e aberto, legível por máquina e com URI para identificação dos dados) [BERNERS-LEE, 2006] ao final do processo.

Como principais contribuições 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, com a utilização de ferramentas de código aberto como o CKAN, Kettle e o *plugin* ETL4LOD.
- 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.

Durante o desenvolvimento do projeto 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.

6. Referências

ARAÚJO L. R.; SOUZA J. F. (2011) “Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados”, http://www.periodicosibepes.org.br/index.php/reinfo/article/view/880, Acesso em: 15 dez. 2020.

BERNERS-LEE, T. (2006) “Linked data”, https://www.w3.org/DesignIssues/LinkedData.html, Acesso em: 9 abr. 2021.

BRASIL (2011) “Lei nº 12.527 de 18 de novembro 2011”, http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm, Acesso em: 10 dez. 2020.

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

GONÇALVES, B. A.; GAMA, K. S. (2018) “Transparência e dados abertos do recife:

uma estratégia bem-sucedida de publicação", https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1901, Acesso em: 8 dez. 2020.

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

OPEN KNOWLEDGE BRASIL (2020) "Por que open", https://www.ok.org.br/dados-abertos/, Acesso em: 11 dez. 2020.

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

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

PENTEADO, B. E.; BITTENCOURT, I. I.; ISOTANI, S. (2019b) "Metaprocesso para transformação de dados educacionais em dados conectados", 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. (2019c) "Modelo de referência para dados abertos educacionais em nível macro", 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. (2021) "Process model with quality control for the production of high quality linked open government data", https://latamt.ieeer9.org/index.php/transactions/article/view/3501, Acesso em: 10 jan. 2021.

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

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