EGCFLOW: UMA APLICAÇÃO DE APOIO AO CICLO DE VIDA DE DADOS ABERTOS CONECTADOS

Jefferson de Oliveira Chaves A[1];
José Leomar Todesco[2].

***Abstract**: The process of producing and maintaining separate open datasets, constituting a complex and costly task, but necessary in the scenario of web data expansion. In view of this, a modeling and implementation of an application was presented to support the production and maintenance of Open Connected Data, through the automation of workflows. An application presented, entitled EGCFlow. As a result, verifying the applicability and application of the application for the contribution of open data was possible, ensuring the process and escalation of repeatability of open data.*

Keywords: Linked Open Data; automation; semantic web; EGCFlow.

Resumo: O processo de produção e manutenção de conjuntos de dados abertos conectados, constitui-se como uma tarefa complexa e onerosa, mas necessária no cenário de expansão da web de dados. Diante disso, apresentou-se a modelagem e implementação de uma aplicação para suporte à produção e a manutenção de Dados Abertos Conectados, por meio da automatização dos fluxos de trabalho, intitulada EGCFlow. Como resultado foi possível verificar a aplicabilidade e contribuição da aplicação para a produção de conjuntos de dados abertos conectados, garantindo escala e repetibilidade a esse processo.

Palavras-chave: Dados Abertos Conectados; automatização; web semântica; EGCFlow.

1. BOAS PRÁTICAS PARA PUBLICAÇÃO DE DADOS E CICLO DE VIDA

Os desafios impostos pela publicação de conjuntos de dados em formatos abertos e conectados, tornaram essencial a adoção de boas práticas para possibilitar que o gerenciamento de conjuntos de dados conectados fosse mais consistente e padronizado. Dentre as boas práticas atualmente recomendadas pela W3C, destacam-se as Boas Práticas para Publicação de Dados na *Web* (DWBP[3], por sua sigla em inglês) propostas por Lóscio, Burle & Calegari (2017) e as Boas

[1] Professor do Instituto Federal de Educação, Ciência e Tecnologia do Paraná (IFPR) – Foz do Iguaçu, Brasil. Mestre em Engenharia e Gestão do Conhecimento, pela Universidade Federal de Santa Catarina (UFSC), – Florianópolis, Brasil. ORCID: https://orcid.org/0000-0002-2476-9100. E-mail: jefferson.chaves@ifpr.edu.br
[2] Professor do Programa de Pós-Graduação Engenharia e Gestão do Conhecimento, pela Universidade Federal de Santa Catarina (UFSC), – Florianópolis, Brasil. ORCID: https://orcid.org/0000-0003-4934-9820. E-mail: tite@egc.ufsc.br
[3] https://www.w3.org/TR/dwbp/.

Práticas para Publicar Dados Conectados propostas por Hyland, Atemezing & Villazón-Terrazas (2014).

As DWBP englobam um universo de variáveis que vão desde o formato de publicação dos dados, até a descrição desses conjuntos de dados e o fornecimento de licenças de uso. A adoção dessas boas práticas, implica em uma série de benefícios, tais como: compreensibilidade; facilidade de processamento; facilidade de descoberta; reúso; confiança; capacidade de conexão de dados; facilidade de acesso; e interoperabilidade (Lóscio et al., 2018).

Uma vez que as DWBP apresentavam boas práticas para publicação de dados de forma geral, e observando o crescente aumento da publicação de dados, principalmente os governamentais, em formato aberto na *web*, o W3C, por meio de um grupo de trabalho criado para o estudo de Dados Conectados Governamentais, estabeleceu uma série de boas práticas para facilitar o desenvolvimento e fornecimento de dados, especificamente, como Dados Abertos Conectados. Essa recomendação ficou denominada como Boas Práticas para Publicação de Dados Conectados (Hyland, Atemezing & Villazón-Terrazas, 2014).

Além da adoção de boas práticas, tornou-se fundamental, para a transformação de dados brutos em Dados Abertos Conectados de alta qualidade e em grande escala, que cada organização adotasse adicionalmente um processo metodológico, que compreenda um conjunto de atividades ou fases bem definidas. A esse processo convencionou-se chamar de Ciclo de Vida de Publicação de Dados Abertos Conectados.

Ressalta-se que os ciclos de vida são diversos e, conforme mencionado, delimitam uma série de fases estabelecidas, que devem ser seguidas para a publicação e consumo de dados na *web*. Neste trabalho, adotou-se a abordagem proposta por Auer (2014), intitulada *Linked Data Lifecycle*.

O *Linked Data Lifecycle*, apresenta um ciclo de vida cujo propósito é o desenvolvimento e compartilhamento de bases de dados conectados na *web*. De acordo com Rautenberg et al. (2018), esse processo foi aplicado com êxito em diversos projetos de publicação de dados conectados, devido, especialmente, ao conjunto de ferramentas computacionais que suportam suas atividades. Ainda, conforme Rautenberg et al. (2018), este ciclo de vida compreende as seguintes etapas: i) Extração; ii) Armazenamento/Consulta; iii) Revisão Manual/Autoria; iv) Interligação/Fusão; v) Classificação/Enriquecimento; vi) Análise de Qualidade; vii) Evolução/Reparação; e viii) Busca/Navegação/Exploração.

As atividades do ciclo de vida proposto no *Linked Data Lifecycle* são suportadas por um ferramental tecnológico denominado de *Linked Data Stack*, cujos objetivos principais são o de facilitar tanto a distribuição, quanto a instalação de ferramentas e componentes, além de diminuir a carga de informações entre os componentes, visando assim uma melhor experiência do usuário (Rautenberg et al., 2018).

O ferramental proposto pela *Linked Data Stack* inspirou a modelagem e a implementação dos algoritmos da aplicação proposta por este trabalho, contudo, optou-se pela sua não utilização, uma vez que, a aplicação proposta tem como um de seus princípios ser independente de domínio, de configurações de ambiente e de orquestração de ferramentas de terceiros.

Para que uma aplicação voltada à *web* semântica seja implementada, contemplando as etapas previstas nos ciclos de vida de publicação de Dados Abertos Conectados, tais como a extração, transformação, mapeamento e conexão com outros conjuntos de dados, e observando outros aspectos fundamentais, como proveniência, reprodutibilidade e repetibilidade de fluxos de execução, foi proposta por Rautenberg et al. (2016) a ontologia denominada *Linked Data Workflow Project Ontology* (LDWPO).

A LDWPO, foi apresentada como um modelo de conhecimento para a estrutura de fluxo de trabalho para Dados Conectados, objetivando auxiliar a descrição do processo metodológico, do plano de fluxo de trabalho e das execuções dos fluxos de trabalho para produção e reprodução de Dados Conectados, possibilitando assim, a reprodutibilidade e repetibilidade das etapas de processamento de Dados Conectados (Rautenberg et al., 2016).

Nesse sentido, a proposta deste trabalho é apresentar uma aplicação para apoiar e dar escala à produção e manutenção de conjuntos de Dados Abertos Conectados por meio da automatização desse processo, apoiando-se nas boas práticas, tendo em vista o ciclo de vida de publicação de dados abertos conectados e tendo como base ontológica a LDWPO. A aplicação em questão é denominada EGCFLOW.

2. DESENVOLVIMENTO DA FERRAMENTA

Assim como já discutido anteriormente, a produção e publicação de Dados Abertos Conectados é uma tarefa complexa. Esse contexto fez com que fossem propostos métodos, técnicas e ferramentas para suporte à execução desta tarefa. Contudo, a orquestração de todo o aparato

tecnológico veio acompanhada de grandes desafios, alguns já citados anteriormente, dentre os quais destacam-se: i) aplicações construídas com caráter *ad hoc*, atendendo a produção de conjuntos de dados específicos; ii) exigência de conhecimento técnico e domínio de distintos *softwares*, que suportem as diversas etapas necessárias para publicação do conjunto de dados em formato aberto e conectado; iii) dificuldade na repetibilidade e reprodutibilidade quanto à produção de conjuntos de dados, sem a necessidade de reconfiguração do ambiente de produção.

Esse cenário motivou o desenvolvimento da ferramenta, intitulada de *Easy*, *General and Creative Flow* – EGCFlow. A ferramenta proposta tem como objetivo, a sistematização do processo de produção de dados abertos e conectados, por meio da criação de fluxos de trabalho, em que conjuntos de dados primários são carregados, transformados, semantificados e conectados a outras fontes de dados, automatizando e dando escala à produção de conjuntos de dados no modelo de metadados RDF.

Para alcançar esse objetivo, o processo de desenvolvimento da ferramenta pautou-se nos princípios da Engenharia do Conhecimento, como fundamento para desenvolvimento de projetos de Gestão de Conhecimento, tais como sistemas de conhecimento e aplicações para processamento e busca de informações.

Ainda, do ponto de vista de Engenharia de *Software*, buscou-se definir o processo para implementação da ferramenta. Além disso, embora o contexto e os requisitos iniciais fossem razoavelmente bem definidos, alguns aspectos não eram bem claros ou mesmo conhecidos. Isso levou a adoção de um modelo de processo iterativo e incremental de desenvolvimento. Esse modelo foi composto das seguintes etapas metodológicas: definição do escopo, levantamento de requisitos, modelagem e implementação. Cada etapa metodológica foi realizada por meio da execução de atividades fundamentadas nos princípios de Engenharia de *Software*, nos princípios estruturais e nas Boas Práticas para publicação de Dados Abertos Conectados.

Todo o processo de desenvolvimento preocupou-se ainda, em proporcionar ao usuário final, fácil instalação e uso. A abordagem adotada permitiu o mapeamento semântico orientado pelo usuário, possibilitando que grandes quantidades de dados pudessem ser convertidas para o modelo de dados RDF, até mesmo por usuários casuais.

A interface implementada, guarda semelhanças comuns a muitos programas de gerenciamento de planilhas, além de não ser necessária a configuração de diversas ferramentas ou

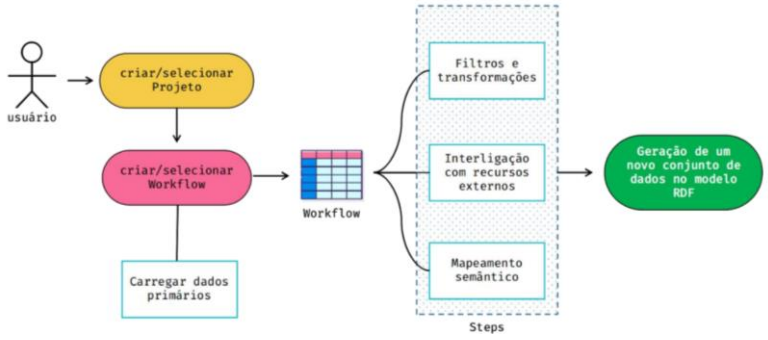
arquivos externos, tendo como únicas dependências o *Java Runtime Enviroment*[4], em sua versão 8 ou superior; o sistema gerenciador de banco de dados MySQL, em sua versão 5.7 ou superior; e um navegador *web*.

A modelagem realizada foi inspirada e construída: i) observando-se as etapas necessárias à produção de dados em um formato aberto e conectado a outros recursos externos, previstas no Ciclo de Vida para Publicação de Dados Abertos Conectados; ii) observando-se as classes e suas relações, propostas na ontologia LDWPO; e iii) observando-se e realizando-se a experimentação da definição de fluxos de trabalho, bem como as atividades e etapas realizadas pelo LODFlow.

Na abordagem apresentada, o usuário deve criar ou selecionar um Projeto existente. O objetivo do Projeto é manter um conjunto de dados conectados, a partir do carregamento de dados primários. Um Projeto pode estar associado a um ou mais *Workflow*s. Um *Workflow* por sua vez, está associado, principalmente, a um conjunto de dados primários e a um conjunto de *Steps*. Um *Step*, assim como na LDWPO, representa uma operação sobre o conjunto de dados primários, com o intuito de transformar, limpar, conectar a outros recursos e semantificar os recursos presentes nesse conjunto de dados. Após a execução de um ou mais *Steps*, o usuário poderá exportar o conjunto de dados resultante para um novo conjunto de dados de acordo com o modelo RDF.

A definição do escopo possibilitou que fosse delineado o fluxo de uso da aplicação. Esse fluxo, simplificado, pode ser observado na Figura 1.

Figura 1 – Fluxo simplificado da aplicação



Fonte: Elaborado pelo autor.

Dentre os principais recursos da aplicação destacam-se: i) a criação de Projetos; ii) a criação de *Workflow*s; iii) seleção de conjuntos de dados iv) a criação de passos para transformação, filtro,

[4] https://www.oracle.com/br/java/technologies/javase-jre8-downloads.html

limpeza, correção dos dados, aqui chamados de *Steps*; v) a interligação desses dados com a DBpedia; vi) o mapeamento semântico e; vii) exportação desses dados.

2.1. *STEPS*

A criação de um *Step*, baseado no conceito LDW*Steps*, da LDWPO tem como objetivo a execução de uma ou mais operações sobre o conjunto de dados de entrada, utilizando-se algoritmos em Java, consultas a banco de dados de triplas ou relacionais, além de outras operações a fim de produzir um conjunto de dados seguindo o modelo RDF.

A implementação dos *Steps* foi desenvolvida de modo a permitir o encapsulamento de parâmetros necessários para realizar a execução, a reversão ou o agendamento de uma ação para execução automatizada.

Essa estratégia possibilitou que a execução, inclusão e reaproveitamento de *Steps* fossem simplificadas. Isso significa que, ao se realizar um *Step* em um *Workflow*, esse mesmo *Step* pode, eventualmente, ser executado mais de uma vez ou mesmo ser utilizado por outros *Workflow*s. A abordagem adotada, ainda permitiu que praticamente qualquer operação pudesse ser implementada como um *Step*, de forma a usufruir os benefícios mencionados acima.

2.2. INTERLIGAÇÃO COM A DBPEDIA

O ciclo de vida proposto por Auer (2014), tem como uma de suas etapas a Interligação ou Fusão. Nesta etapa, um conjunto de dados é interligado a outro, com o intuito de ampliar os contextos de pesquisa e consulta.

Assim, este *Step* teve como objetivo vincular os recursos presentes no conjunto de dados primários a outros conjuntos de dados (*mashup* semântico), nesse caso a DBpedia[5], de modo que a união entre esses conjuntos de dados contribuísse para criação de dados mais completos. Além da interligação entre conceitos ser um dos princípios da *web* semântica e viabilizar a chamada "*Web* de Dados Conectados", a interligação faz com que o conjunto de dados resultado alcance cinco estrelas, de acordo com o modelo proposto por Berners-Lee.

[5] A DBpedia é um dos principais nós da *web* de dados. Foi utilizada para este projeto como ponto de conexão para dados abertos conectados.

A realização desse *Step* faz com que cada linha da coluna selecionada seja associada a um recurso presente na DBpedia, quando existir, desde que a informação presente nessa linha contenha o mesmo nome ou identificador desse recurso, e que os tipos selecionados pelo usuário estejam presentes nesse mesmo recurso.

Os resultados encontrados irão compor as triplas do modelo RDF por meio do predicado <owl:sameAs>, que significa que dois URIs diferentes representam o mesmo recurso do mundo real, ligando os recursos do conjunto de dados com os da DBpedia. Ao ser salvo, esse *Step* será agendado para ser executado no momento de exportação do conjunto de dados RDF.

2.3. MAPEAMENTO SEMÂNTICO

O objetivo deste *Step* é criar um modelo semântico para definição dos recursos existentes no domínio, por meio da atribuição de propriedades e relacionamentos que podem ser usados para descrevê-los, apoiando-se pelo uso de vocabulários e ontologias de referência. Esse modelo deve ser projetado pelo usuário, de modo a selecionar os dados que irão compor as triplas que farão parte do modelo RDF resultante.

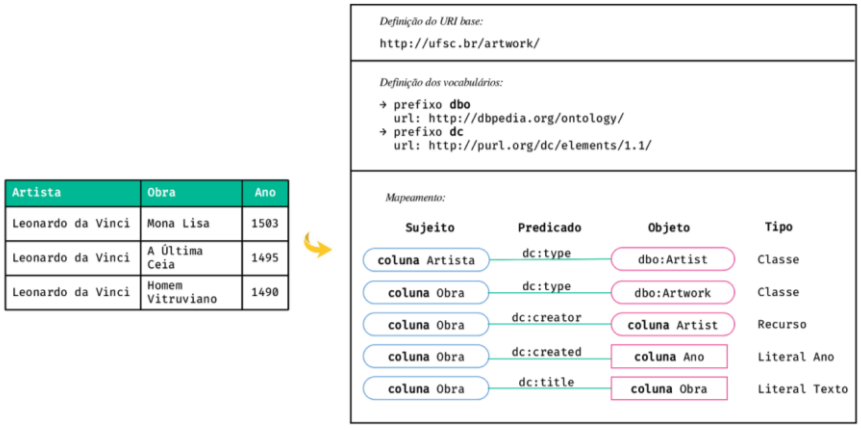
A abordagem utilizada consistiu na criação de um modelo baseado em triplas RDF, para o mapeamento das informações presentes na representação tabular, de modo que as colunas, identificadas por seus cabeçalhos, fossem descritas como parte de uma tripla RDF, na forma <sujeito><predicado><objeto>.

Os sujeitos são definidos por meio da seleção da coluna da representação tabular, ao qual se deseja atribuir predicados. Os sujeitos devem ser recursos identificados por URIs. Os predicados especificam como os sujeitos e os objetos estão relacionados e devem ser definidos por meio do uso de vocabulários de referência que melhor identifiquem o conjunto de dados, aumentando a expressividade e reduzindo ambiguidades. Este *Step* oferece suporte a reutilização de ontologias e vocabulários existentes. Dessa forma, o vocabulário deve ser importado, informando-se um prefixo e o *link* para o vocabulário, ou realizando o *upload* do arquivo, nos formatos RDF e OWL, com o vocabulário. Ao se importar um vocabulário, são importadas suas propriedades (propriedades de objeto e de tipo de dados) e suas classes, que podem ser usadas para representar um objeto na tripla. Por fim, os objetos podem ser representados por literais definidos pelo XML *Schema DataType*,

por classes da ontologia importada ou por outros recursos, selecionando-se as colunas ao qual se deseja conectar.

Quando um recurso é mapeado, é necessário que esteja devidamente identificado por URIs únicos, uma vez que representa um recurso específico na *web* de dados. Nesse sentido, este *Step* gerencia a escrita automatizada de URIs de acordo com o modelo {*workflow*_uri}/{column}/{data}. Nesse modelo, {*workflow*_uri} está relacionada com um URI válida, definida no momento de criação do *Workflow*. O {column}, por sua vez, refere-se à coluna em que o recurso está disposto e {data} descreve à própria coisa a qual se refere o recurso. A Figura 2 ilustra o procedimento para realizar o mapeamento semântico.

Figura 2 – Mapeamento Semântico

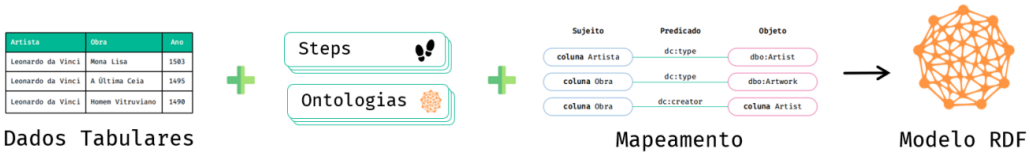


Fonte: Elaborado pelo autor.

Além dos *Steps* de interligação com a DBpedia e de mapeamento semântico, outros *Steps* foram implementados, com o intuito de realizar transformações no conjunto de dados primários.

É importante salientar que, os *Steps* executados em um *Workflow*, podem ser reaproveitados em outros *Workflow*s. Para concretizar a criação da base de conhecimento ainda é necessária a execução de uma última etapa pelo usuário, chamada aqui de exportação. É nesta etapa que os dados dispostos no modelo tabular são traduzidos para o modelo RDF. A Figura 3 representa um esquema desta etapa.

Figura 3 – Visão geral do processo de geração do Modelo RDF



Fonte: Elaborado pelo autor.

A execução desta etapa é gerenciada por um *WorkflowExecution*, responsável pela criação da base de conhecimento, por meio da combinação do *Step* de Mapeamento Semântico e dos dados no modelo tabular. O *WorkflowExecution* ainda é responsável por verificar os *Steps* agendados para antes ou após a criação da base de conhecimento e executá-los.

3. CASO DE USO DO EGCFLOW: ELEIÇÕES 2018

A fim de demonstrar seu uso, optou-se pela utilização de dados primários do mundo real, referentes às eleições no Brasil. A complexidade inerente ao processo eleitoral, associado ao desafio gerado pelo fato de que os dados relacionados às eleições, muitas vezes, não estão adequados ao processamento computacional, acaba por gerar inconvenientes em sua coleta, modelagem, padronização, consumo e combinação com outros conjuntos de dados, impactando no uso dos mesmos e, consequentemente, na produção de conhecimento acerca deles.

3.1. CRIAÇÃO DE PROJETO E *WORKFLOW*S

Para criação do projeto, foram registrados: i) nome do projeto: “Eleições 2018”; ii) homepage: <http://ufsc.edu.br/eleicoes>; iii) autor: John Doe; iv) palavras-chave: Eleições, Eleições Federais, Eleições 2018; v) idioma: pt-br; vi) propósito: transformar e conectar dados das eleições de 2018; v) descrição: Número de votos por candidato das eleições de 2018. É importante destacar que após a criação do projeto, automaticamente, foi criada uma pasta no computador do usuário, em que foram salvos os artefatos, separados por *Workflow*, referentes a este Projeto.

Um Projeto é o ponto de partida para criação de *Workflow*s. Cada *Workflow* de um Projeto possui um conjunto de dados primários relacionados a um mesmo conteúdo ou tema, mas que

apresenta distinções quanto a sua categoria, grupo, licença, cobertura geográfica ou tempo. Neste exemplo, os *Workflow*s foram criados baseados na unidade federativa em que os dados foram produzidos, ou seja, por cobertura geográfica. A seleção dos estados ocorreu por ordem alfabética, de forma que o estado do Acre foi o primeiro selecionado. Para a criação de um *Workflow*, é necessário informar: i) um nome; ii) um URI, que será utilizado como base para criação dos recursos; iii) fonte primária; iv) cobertura espacial; v) período temporal; vi) o padrão de codificação do documento; vii) o caractere separador, para o caso de arquivos CSV ou TSV; viii) uma licença; e por fim, ix) o próprio arquivo contendo os dados primários.

O *Workflow* em questão foi intitulado de "Votação Acre". Seu URI foi definido como <http://ufsc.br/eleicoes/votacao_acre/2018/v1/>. Esse URI permite que seja identificada a versão desse conjunto de dados (v1), além de compor os URIs dos recursos contidos no conjunto de dados. Definiu-se a fonte primária dos dados como o *link* que aponta para sua fonte original, qual seja <https://cdn.tse.jus.br/estatistica/sead/eleicoes/eleicoes2018/bu*web*/B*WEB*_1t_AC_10102081938.zip>.

A cobertura espacial, por sua vez, foi definida como o estado do Acre. Já o período temporal foi definido como 2018. A licença foi selecionada, com base na Portaria nº 93, de 12 de fevereiro de 2021, do TSE, que define os dados em questão como dados abertos. Assim a *The Open Government License* (OGL) foi selecionada, dentro de um rol de 10 (dez) licenças disponibilizadas pela aplicação. O padrão de codificação utilizado, segundo informado pelo TSE no arquivo "leiame", é o Latin 1, também conhecido como "ISO_8859_1", sendo, portanto, este padrão o selecionado para o *Workflow*. O caractere separador "ponto e vírgula" foi selecionado para este conjunto de dados e, por fim, o arquivo contendo os dados do Acre, no formato CSV, foi selecionado.

Ao ser criado um *Workflow*, uma pasta com seu nome é criada dentro da pasta do Projeto e o conjunto de dados primários é armazenado nesta pasta. Concluída a criação do *Workflow*, é possível a execução dos *Steps* para transformar, semantificar e gerar os dados no modelo RDF, etapa descrita no próximo item.

3.2. EXECUÇÃO DE *STEPS*

Após a criação do projeto e do *Workflow*, foram executados uma série de *Steps*, a fim de melhor estruturar os dados. As alterações no conjunto de dados realizado por cada *Step* podem ser vistas abaixo: i) renomear cabeçalhos: neste *Step*, os cabeçalhos das colunas da tabela de dados foram renomeados. Colunas com abreviações foram renomeadas com seus nomes por extenso, a fim de melhor descrever os recursos associados; ii) reordenar Conjunto de Dados: neste *Step*, o conjunto de dados foi reordenado, a partir da execução do *Step* "Reordenar Conjunto de Dados", aplicado à coluna "MUNICÍPIO". Cabe destacar que, a reordenação, ainda que tenha sido aplicada à coluna supracitada, se estende às demais colunas do conjunto de dados. Dessa forma, a tabela de dados foi disposta de maneira organizada, em ordem alfabética crescente; iii) normalizar caixa alta/caixa baixa: os dados primários apresentavam, ora dados em caixa baixa, ora dados em caixa alta. Assim optou-se, para fins de padronização, a transformação das colunas com textos em caixa baixa para caixa alta; iv) filtrar: o *Step* de Filtro, permite que linhas sejam filtradas de acordo com o valor de seus campos. Neste caso, foi aplicado o *Step* de Filtro à coluna "DESCRIÇÃO TIPO VOTÁVEL", com o objetivo de manter no conjunto de dados, apenas os dados referentes a votos nominais, excluindo linhas cujos valores dessa coluna fossem iguais a "LEGENDA", "BRANCO" ou "NULO".

3.3. EXECUÇÃO DO MAPEAMENTO SEMÂNTICO

Neste *Step*, foi necessário realizar manualmente um conjunto de tarefas para o mapeamento. A primeira delas consistiu na importação das ontologias e vocabulários para definição das classes e propriedades, que iriam compor o mapeamento. Deste modo, foram importadas a ontologia de "Eleições", identificada pelo prefixo <ele>, o vocabulário RDF, identificado pelo prefixo <rdf> e o vocabulário *Dublin Core*, identificado pelo prefixo <dc>. Após a configuração das ontologias e vocabulários, é necessário realizar o mapeamento semântico. O mapeamento segue o modelo proposto pelas triplas RDF, estabelecendo uma relação binária entre dois recursos, ou um recurso e um literal, conectados por meio de um predicado.

3.4. INTERLIGAR À DBPEDIA

Neste *Step*, os recursos da coluna "MUNICÍPIO" foram associados a recursos presentes na DBPedia, quando existentes. Para execução deste *Step* foi necessário selecionar as classes que melhor descreviam os recursos dessa coluna. As classes listadas foram obtidas a partir de uma amostra, neste caso, uma cidade presente na coluna.

Como resultado, foram retornados 65 (sessenta e cinco) recursos, cujo rótulo ou nome continham o texto "Rio Branco". Entre esses recursos, figuravam recursos relacionados a rios, times de futebol, lugares, pessoas e cidades. Assim, com o objetivo de filtrar tais resultados foi selecionada, dentre as classes disponibilizadas, a classe <http://dbpedia.org/ontology/Town>, que representa o conceito de cidade. Como resultado foram retornados quatro recursos, quais sejam: Rio Branco do Ivaí; Rio Branco, Mato Grosso; Rio Branco, Acre; e Rio Branco do Sul.

Como a classe selecionada se aplicava a outros recursos, que não somente a cidade Rio Branco para o Estado do Acre, optou-se pela seleção de uma classe mais específica, a classe <http://dbpedia.org/class/yago/WikicatPopulatedPlacesInAcre(state)>, que retornou de forma correta os recursos relacionados à cidade do estado do Acre.

Foi realizada uma verificação manual de cada uma das 22 (vinte e duas) cidades do estado, a fim de averiguar a consistência do *Step*. Realizada a seleção da classe para interligação com a DBpedia, foi necessário salvar o *Step*.

3.5. EXPORTAÇÃO DE DADOS

Realizados os *Steps* desejados, foi necessário realizar a exportação dos dados do *Workflow*. Suportado pelo mapeamento semântico, a exportação dos dados criou os modelos RDF serializados nos formatos *Turtle*, JSON-LD e RDF/XML.

A fim de fornecer informações e proveniência acerca do conjunto de dados gerado, nesta etapa foi criado um relatório em formato RDF. Neste relatório foram adicionados metadados descritivos e estruturais, com o objetivo de enriquecer o entendimento, tanto do conjunto de dados, quanto do seu processo de criação. Além disso, foram adicionadas informações acerca da proveniência da fonte primária de dados. Nesse sentido, o relatório criado apresenta as informações de título, autor, origem, palavras chave, data de criação, cobertura espacial, período temporal,

licença, objetivo, descrição, bem como a lista de *Steps* executados para transformar, interligar e gerar o conjunto de dados em RDF.

4. CONSIDERAÇÕES FINAIS

Este trabalho se propôs a pensar soluções orientadas para sanar o problema de apoio e escalabilidade à produção e manutenção de Dados Abertos Conectados, por meio da automatização de fluxos de trabalho. É necessário que os dados sejam publicados seguindo práticas, que permitam que os dados possam ser compreendidos e utilizados por consumidores, além de ser passíveis de processamento por agentes de *software*.

A utilização do EGCFlow por meio do emprego em cenário real das eleições brasileiras, demonstrou a aplicabilidade da aplicação e o cumprimento do objetivo proposto para o trabalho: apoiar e dar escala à produção e manutenção de conjunto de Dados Abertos Conectados, por meio da automatização de fluxos de trabalho.

A aplicação desenvolvida resultou em uma série de benefícios de ordem técnica, dentre os quais destacam-se: i) a produção de conjuntos de dados abertos conectados, sem a dependência de distintas ferramentas ou conhecimentos técnicos aprofundados; ii) a flexibilidade de utilização, possibilitando o uso para distintos casos, uma vez que o do sistema eleitoral brasileiro utilizado aqui, foi apenas um exemplo; iii) a criação e reutilização dos fluxos de trabalho, otimizando, dentre outros fatores, o tempo; e iv) a interface semiautomatizada para execução dos fluxos de trabalho. De forma indireta a aplicação tem potencial para proporcionar outros benefícios quando adotada. Ao se analisar o que é possível realizar, para além da execução da aplicação, mas observando-se os benefícios de se utilizar os conjuntos de dados por ela produzidos, é possível destacar diversos benefícios relacionados à gestão para organizações de caráter tanto público quanto privado, dentre os quais destacam-se: i) aumento do uso e reúso de dados; ii) aumento na transparência; iii) geração de valor; iv) *accountability*; v) aumento na qualidade de dados; vi) apoio à tomada de decisões; e vii) geração de conhecimento. Todos estes elementos contribuem, direta ou indiretamente, nos processos de gestão do conhecimento em distintas organizações.

Assim, considerando-se as múltiplas iniciativas de caráter internacional, como a *Open Government Partnership*; a previsão do direito de acesso à informação na Declaração Universal dos Direitos Humanos, da qual o Brasil é signatário, bem como na própria Constituição Federal de

1988; a Lei de Acesso à Informação e o Decreto que institui a Política de Dados Abertos do Poder Executivo Federal, é possível observar que há cada vez mais um movimento na direção para a expansão de uma *Web* de Dados.

REFERÊNCIAS

Auer, Sören. 2014. Introduction to LOD2. Chapter Linked Open Data Creating Knowledge Out of Inter*Linked Data*. Volume 8661 of the series LEcture Notes in Computer Science. p. 117.

Hyland, Bernadette; Atemezing, Ghislain; Villazón-Terrazas, Boris. 2014. Best Practices for Publishing *Linked Data*. W3C Working Group. Note 09. Disponível em: < https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook>. Acesso em: 22 de Mai. de 2022.

Lóscio, Bernadette; Burle, Caroline; Calegari, Newton. 2017. Data on the *Web* Best Practices. W3C Recommendation. Disponível em: < https://www.w3.org/TR/dwbp/#bib-LD-BP>. Acesso em: 28 de Mai. de 2022.

Lóscio, Bernadette et al. 2018. Fundamentos para publicação de dados na *web*. São Paulo: Comitê Gestor da Internet no Brasil.

Rautenberg, Sandro, et al. 2018. Guia Prático para Publicação de Dados Abertos Conectados na *Web*. Curitiba: Appris.

Rautenberg, Sandro, et al. 2016. LDWPO – A lightweight *Ontology* for *Linked Data* management. CEUR Workshop Proceedings.