DBacademic: Conectando os dados abertos das instituições de ensino do Brasil

Sérgio Souza Costa
Doutor em Computação Aplicada pelo Instituto Nacional de Pesquisas Espaciais (INPE) - São José dos Campos, SP – Brasil. Professor da Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/2073311645132958
E-mail: sergio.costa@ufma.br

Mateus Vitor Duarte Sousa
Graduando em Bacharelado Interdisciplinar em Ciência e Tecnologia pela Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/7602019586262918
E-mail: mateusriograndense@gmail.com

Micael Lopes da Silva
Graduado em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA) - São Luís, MA -Brasil.
http://lattes.cnpq.br/3876640219126946
E-mail: micaelopes32@gmail.com

Eddye Cândido de Oliveira
Graduando em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA) - São Luís, MA - Brasil
http://lattes.cnpq.br/8610296132900395
E-mail: eddyeoliver@gmail.com

José Victor Meireles Guimarães
Graduando em Engenharia da Computação pela Universidade Federal do Maranhão (UFMA), São Luís, MA – Brasil.
http://lattes.cnpq.br/7974690550998329
E-mail: jvictormguimaraes@gmail.com

Submetido em: 24/11/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

As instituições públicas detêm um grande volume de dados que poderiam ser usados para melhorarem os seus serviços. Isso motivou um movimento denominado de dados abertos. Neste sentido, o Brasil e outros países, têm criado leis que incentivam e, de maneira compulsória, garantem que as instituições abram os seus dados públicos. Por meio do Decreto nº 8.777 de 2016 foi definido que todas as instituições federais deveriam elaborar o seu Plano de Dados Abertos (PDA). Esse decreto levou à criação e publicação de um grande volume de dados abertos pelas diversas instituições públicas. Atualmente, cada instituição mantém isoladamente os seus dados, o que torna impraticável a consulta de dados entre elas. O objetivo deste trabalho é, então, conectar esses dados em um grande repositório de dados denominado DBacademic. Para isso, dados abertos de 25 instituições públicas de ensino foram extraídos, transformados e carregados nesse repositório. Essa transformação resultou em quase 900 mil triplas que podem ser consultadas no endereço www.dbacademic.tech. Os resultados mostram o potencial dessa solução para possibilitar diversas consultas relevantes que seriam muito difíceis de serem realizadas com os repositórios isolados.

Palavras-chave: Dados Conectados. Universidades. RDF. PDA. Repositório.

DBacademic: Linking the open data of educational institutions in Brazil

ABSTRACT

Public institutions have a large amount of data that could be used to improve their services. This has motivated a movement called open data. In this sense, Brazil and other countries have created laws that encourage and, in a compulsory manner, guarantee that institutions open their public data. Through the Decree nº 8.777 of 2016, it was defined that all federal institutions prepared their Open Data Plan. This decree led to the creation and publication of a large volume of open data by the various public institutions. Currently, each institution maintains it's data in isolation, which makes it impossible to query data between them. The purpose of this work is then to connect these data to a large data repository called DBacademic. To that end, open data from 25 public educational institutions were extracted, transformed and uploaded to this repository. This transformation resulted in almost 900 thousand triples that can be queried through the following address: www.dbacademic.tech. The results showed the potential of this solution to enable several relevant queries that would be very difficult to be carried out as isolated repositories.

Keywords*: Open data. Universities. RDF. Open Data Plan. Repositories.*

DBacademic: Vinculando los datos abiertos de las instituciones educativas en Brasil

RESUMEN

Las instituciones públicas tienen una gran cantidad de datos que podrían utilizarse para mejorar sus servicios. Esto ha motivado un movimiento llamado datos abiertos. En este sentido, Brasil y otros países han creado leyes que fomentan y, de manera obligatoria, garantizan que las instituciones abran sus datos públicos. Mediante el Decreto nº 8.777 de 2016, se definió que todas las instituciones federales elaboraron su Plan de Datos Abiertos. Este decreto propició la creación y publicación de un gran volumen de datos abiertos por parte de las distintas instituciones públicas. Actualmente, cada institución mantiene sus datos de forma aislada, lo que hace que sea imposible consultar datos entre ellas. El propósito de este trabajo es entonces conectar estos datos a un gran repositorio de datos llamado DBacademic. Para ello, se extrajeron, transformaron y subieron a este repositorio datos abiertos de 25 instituciones educativas públicas. Esta transformación resultó en casi 900 mil triples que se pueden consultar a través de la siguiente dirección: www.dbacademic.tech. Los resultados mostraron el potencial de esta solución para permitir varias consultas relevantes que serían muy difíciles de realizar como repositorios aislados.

Palabras clave: *datos abiertos. Universidades. RDF. Plan de datos abiertos. Repositorios.*

1 – INTRODUÇÃO

No Brasil, o acesso aos dados de instituições públicas já estava previsto pela Constituição de 1988, porém foi reforçado por meio da Lei de Acesso à Informação (Lei n.º 12.527/2011). Adicionalmente, o Decreto 8.777, de maio de 2016, definiu que órgãos e entidades da administração pública federal deveriam elaborar e publicar um Plano de Dados Abertos (PDA). Em resposta a essa demanda, as instituições disponibilizam um grande volume de dados públicos e abertos em seus portais. Esses dados propiciaram uma maior participação da comunidade no desenvolvimento de soluções inovadoras que melhoram e fiscalizam os serviços públicos. Dentre essas instituições federais, este artigo focalizou as de ensino superior, técnico e tecnológico, como as Universidades Federais e os Institutos Federais de Ciência e Tecnologia (IFETs). Na literatura, é possível encontrar alguns trabalhos que destacam a demanda de acesso aos dados dessas instituições, como Carossi e Teixeira Filho (2017), Gama e Rodrigues (2016), Zorzal e Rodrigues (2016).

Muitas destas instituições de ensino possuem portais específicos para o acesso a seus dados abertos. Neste trabalho, foram identificadas 45 instituições, com uma média de 20 conjuntos de dados cada. Esses conjuntos de dados incluem informações sobre servidores, estudantes, projetos de pesquisa, despesas e orçamentos. Eles são indexados pelo Portal Brasileiro de Dados Abertos (http://www.dados.gov.br/) para facilitar a busca. Mesmo sendo indexados por este portal, tais dados permanecem isolados e de difícil integração. oAssim, diversas perguntas relevantes são difíceis de serem respondidas, como: Quais são os nomes (ou endereços eletrônicos) dos coordenadores dos cursos de Engenharia da Computação das instituições públicas de ensino do Brasil?

Mesmo com esses nomes presentes na maioria dos portais de dados abertos, esta pesquisa iria requerer um algoritmo específico para consultar cada um dos portais em busca dessa informação. Esse algoritmo teria que lidar ainda com os diferentes modelos de dados das instituições, tornando sua escrita impraticável, poi, a consulta manual nos portais Web seria provavelmente mais rápida e eficiente.

Esse é um exemplo simples, mas capaz de mostrar como a conexão entre os dados poderia aumentar muito a expressividade das consultas, o que facilitaria a análise de dados e o desenvolvimento de soluções inovadoras. Essa conectividade é possível a partir de um conjunto de boas práticas propostas por Tim Berners-Lee, denominadas Dados Conectados. Existem, atualmente, diversos repositórios de dados conectados, como pode ser observado no diagrama da figura 1. Nele, também é possível observar que o DBpedia é, provavelmente, o repositório mais conhecido e conectado (AUER *et al.*, 2007). No contexto de dados de instituições de ensino, alguns trabalhos têm proposto repositórios. Assim sendo, foram considerados, neste artigo, os estudos de Alencar *et al.* (2018), Costa *et al.*(2019), Kessler e Kauppinen (2015), Piedra *et al.*(2014), Rocha e Lóscio (2015) e Zablith, Fernandez e Rowe (2012).

Alencar et al (2018) aventaram a publicação e o consumo de dados conectados da Unidade Acadêmica de Informática (UAI) do IFPB (Instituto Federal da Paraíba) em conjunto com uma ontologia própria, denominada OpenUAI. Rocha e Lóscio (2015) também propuseram a publicação e a criação de uma ontologia para o Centro de Informática da Universidade Federal de Pernambuco (CIn/UFPE). Costa *et al.* (2019) apresentaram uma metodologia semiautomática para a extração de dados públicos e a sua transformação e carga como dados abertos conectados. Como caso de uso, os autores utilizaram os portais de dados públicos da Universidade Federal do Maranhão, ao invés do portal oficial de dados abertos. Em geral, esses artigos demonstram os ganhos de expressividade alcançados ao conectar os diversos dados de uma dada instituição de ensino.

Durante a pesquisa, muitos dos trabalhos citados não estavam em funcionamento. Além disso,aem sua maioria, eles não tinhammo objetivo de conectar os dados de toda a instituição. Outro desafio para integrá-los seria a necessidade de compatibilizar os seus distintos vocabulários.

Usando o conceito de dados conectados, este artigoovisa a propor um repositório de dados qur irá permitir responder perguntas que integrem dados de diferentes instituições de ensino do Brasil. Perguntas que atualmente seriam impossíveis, ou impraticáveis de serem respondidas. Este repositório, denominado DBAcademic, já agrega atualmente dados de 25 instituições Brasileiras de autarquia federal.

Este artigo está organizado da seguinte maneira: seção 2 apresenta alguns conceitos essenciais, seção 3 apresenta a metodologia do trabalho, seção 4 apresenta os principais resultados e desafios, e seção 5 apresenta as conclusões.

2 – FUNDAMENTOS

Existe uma linha tênue entre os conceitos de dados públicos, abertos e conectados. Nesse sentids, o conceito de dados é o mais básico e estudado pelos cientistas da informação. Semeler e Pinto (2019) apresentam muitas destas definições em um ensaio sobre os dados de pesquisa. Nesse ensaio, os autores consideraram os dados como qualquer objeto criado em formato digital, ou convertido para este, que possa ser usado para geração de *insights* de informação e conhecimento.

Os dados públicos são um subconjunto detses dados, que segundo a Controladoria Geral da União do Brasil, são informações que não lesem leis de privacidade, integridade e segurança (COSTA, *et al.*, 2013). No Brasil, a Lei nº 12.527, de novembro de 2011, regularizou o direito de acesso às informações de órgãos públicos administrativos, autarquias, fundações e empresas estatais (BRASIL, 2020).

Os dados abertos são aqueles que além de poderem ser usados, modificados e compartilhados precisam seguir alguns princípios, como: serem completos, primários, atuais, acessíveis, compreensíveis por máquina, não proprietários e livres de licença (OPEN GOVERNMENT WORKING GROUP, 2007; OPEN KNOWLEDGE FOUNDATION, 2019). Essa definição foi então reforçada pelo art. 2º do Decreto no 8.777/2016:

> III dados abertos-– dados acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte.

Os dados abertos atendem ao critério de transparência proativa, pois o detentor dos dados não espera uma solicitação para disponibilizá-los (ZORZAL; RODRIGUES, 2016). Etsa categoria de transparência deveria ser a forma principal seguida pelo estado para disponibilizar os seus dados, como destacado em (Zorza; e Rodrigues, 2016, p. 2)

> A informação sob a tutela do Estado é um bem público e sua evidenciação deve ser por iniciativa da Administração Pública, de forma espontânea, proativa, independente de qualquer solicitação, ou seja, transparência ativa, como definido em lei.

O Decreto no 8.777/2016 instituiu a política de dados abertos do poder executivo federal brasileiro e foi um importante passo nessa direção. Nesse decreto, definiu-se uma data limite para que órgãos e entidades da administração pública federal elaborassem e publicassem seus Planos de Dados Abertos (PDA). Atender a essa exigência foi, e ainda é, um grande desafio para muitas instituições federais, como destacado em Bertin *et al.* (2017) e Torino, Trevisan e Vidotti (2019). Mesmo com todos os desafios enfrentados pelas instituições, esse decreto ampliou muito a disponibilidade de portais de dados abertos das diversas instituições brasileiras.

Os dados abertos já têm um grande potencial para aumentar a participação da comunidade, melhorando a fiscalização e a qualidade dos serviços prestados pelas instituições públicas. Contudo, realizar consultas integrando estes dados é muito diiícil, sobretudo usando os formatos que são utilizados geralmente pelos portais de dados abertos. Bernes-Lee (2009) propõe então o conceito de dados conectados, chamando a comunidade para construir a "Web dos dados", em contraposto a atual "Web das páginas".

Em resumo, o conceito de dados conectados (ou ligados) refere-se a um conjunto de boas práticas para publicar e ligar os dados estruturados na Web (HEATH; BIZER, 2011). Dentre essas boas práticas, Tim Berners-Lee estabeleceu quatro princípios (BERNERS-LEE, 2009):

1. Use URIs (*Uniform Resource Identifier*) para identificar os recursos;
2. Use HTTP URIs de forma a possibilitar que as pessoas possam procurar esses recursos na Web;
3. Quando alguém procurar por uma URI, forneça informações relevantes utilizando formatos padrões;
4. Inclua conexões para outras URIs de forma a possibilitar que mais recursos possam ser descobertos.

O primeiro princípio exige o uso de *URIs* para identificar os recursos a serem publicadas,Eenquant, o segundo princípio assegura o uso desses identificadores por meio de requisições Web. Já o terceiro princípio estabelece que, quando solicitado um recurso, este deve ser fornecido com todas as suas informações em um formato de dados padrão mantido pela W3C[1], como o *Resource Description Framework* (RDF). Por último, o quarto princípio refere-se à conexão com dados já existentes.

Ao seguir esses princípios, será possível transformar a Web em um banco de dados global, como pode ser observado pelo diagrama criado pelo T*he Linked Open Data Cloud*[2] e ilustrado na figura 1. Nela é possível visualizar a conexão, representads por linhas, entre diversas bases de dados criadas e mantidas por diferentes instituições do mundo.

No centro da nuvem apresentada na figura 1, destaca-se a base de dados DBpedia. Esse projeto coletou os dados da Wikipédia e os disponibilizou no formato de dados conectados (MCCRAE *et al.,* 2020). Segundo McCrae *et al.* (2020), a DBpedia trata a Wikipédia como um banco de dados e tem o objetivo de extrair suas informações estruturadas e torná-las disponíveis na Web para qualquer outra base de dadosP permitindo, assim, incluir estes dados nas suas consultas, como será apresentado em um exemplo na seção 4.

Os dados abertos e conectados são aqueles que atendem a ambas as propriedades. Segundo (RIBEIRO 2015), os repositórios de dados abertos que seguem os princípios de dados conectados são hoje uma alternativa mais consistente e viável do que a disponibilização de documentos, arquivos com metadados. No esquema de classificação, proposto por Tim Berners-Lee (2009), os dados abertos e conectados possuem a avaliação mais alta em um sistema de classificação de 5 estrelas, figura 2.

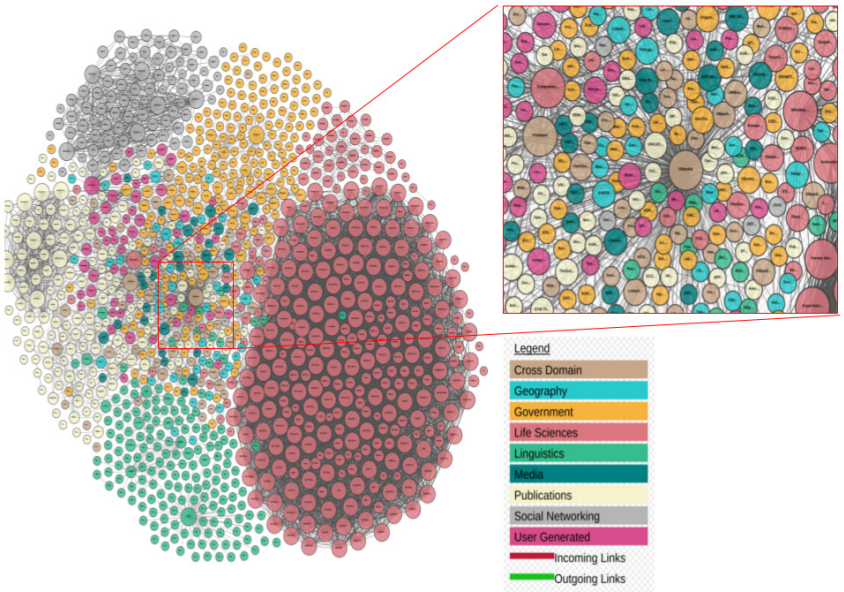
Esse sistema apresentado na figura 2 classifica o nível de abertura de dados. Assim, quanto mais alto o número de estrelas, mais fáceis esses dados estão de serem conectados. Cada estrela é atribuída a um nível de abertura dos dados, partindo da disponibilidade dos dados de maneira pública até chegar àqueles que estão conectados a outras bases. São eles:

1. Abastecimento dos dados públicos na Internet com licença aberta;
2. Utilização de formatos estruturados ao invés de páginas HTML;
3. Utilização de formatos não proprietários, como o CSV;
4. Emprego de padrões estabelecidos pela W3C, como o RDF;
5. Inclusão de conexões a outros dados já existentes.

[1] A W3C (www.w3.org) é a principal organização de padronização da World Wide Web.

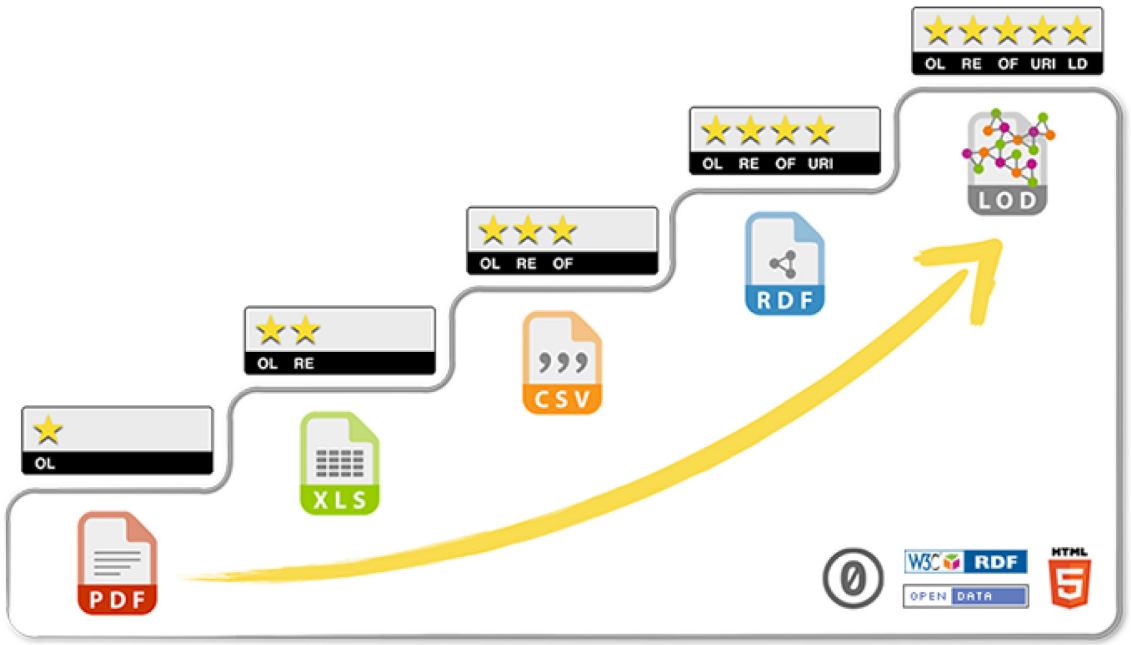
[2] O diagrama original pode ser acessado em https://lod-cloud.net/

Figura 1 – Diagrama de uma nuvem LOD em 2017



Fonte: Adaptado de McCrae *et al.* (2020).

Figura 2 – Classificação cinco estrelas



Fonte: Berners-Lee (2009).

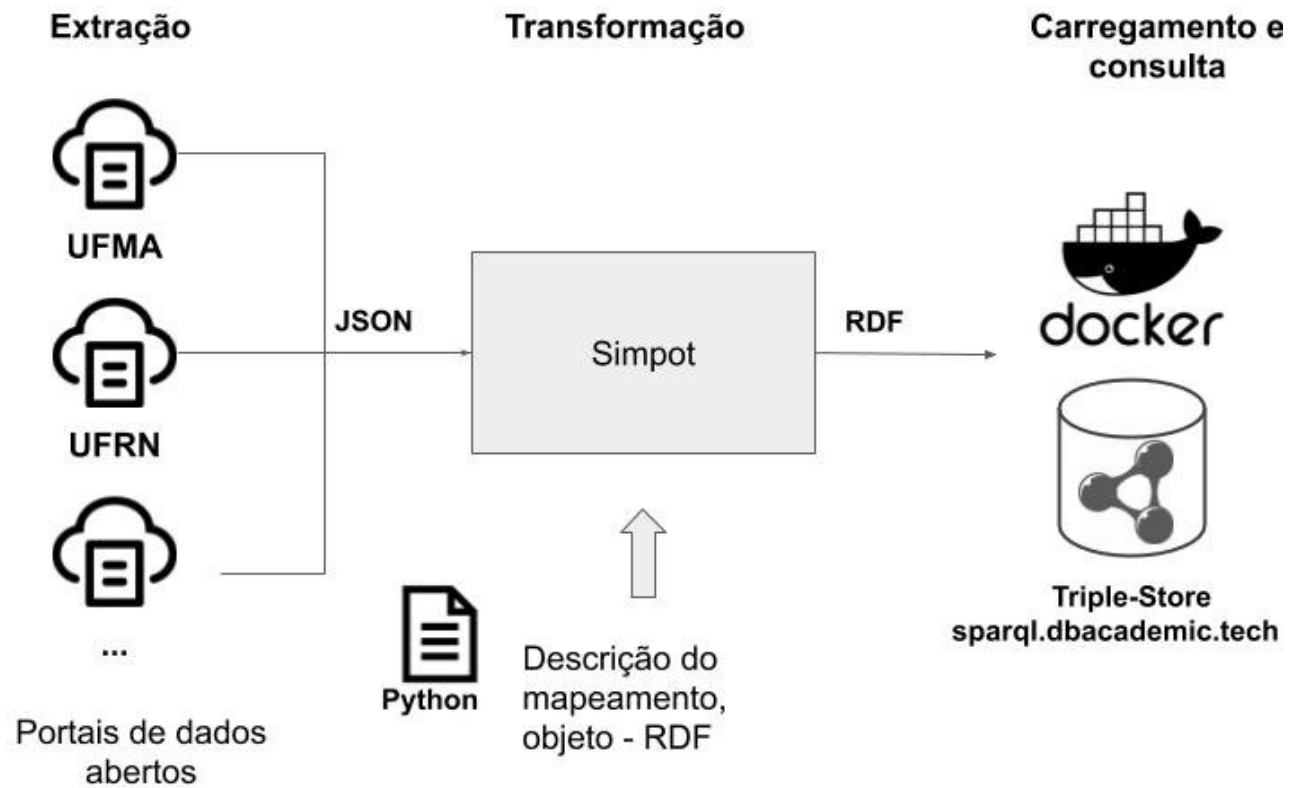
A primeira estrela é adquirida quando os dados são publicados sob licença aberta de uso. No entanto, características como a estruturação dos dados não sãé ainda algo fundamental. A segunda estrela é alcançada quando os dados já se encontram estruturados em formatos proprietários como o XLS da Microsoft. Adicionalmente, para receber a terceira estrela estes dados precisam estar em algum formato aberto, podendo ser CSV ou JSON. Para a próxima estrela já é necessário utilizar o formato de dados RD(. Esse formato de dados foi proposto pela W3C no contexto da Web Semântica e permite associar os recursos a conceitos de um ou mais vocabulários. Por fim, para atingir a quinta estrela, é necessário que os recursos se conectem, sempre que possível, a dados e vocabulários já existentes.

De acordo com esta classificação, os portais de dados abertos das instituições de ensino estariam classificados com 3 estrelas, enquanto o repositório DBacademic tem como objetivo atender todos os cinco princípios.

3 – METODOLOGIA

Para alcançar os objetivos propostos neste artigo, utilizou-se uma adaptação da abordagem semiautomática para a extração, a transformação e o carregamento de dados conectados proposto a em Costa *et al.* (2019). Essa metodologia foi influenciada pelo conceito *Extract Transform Load* (ETL), que é usualmente utilizado no contexto de Data Warehouse (VASSILIADIS; SIMITSIS; SKIADOPOULOS, 2002). Uma visão geral da metodologia é apresentada na figura 3.

Figura 3 − Visão geral da metodologia



Fonte: Elaboração do autor (2020).

Este trabalho difere, principalmente, em dois pontos àa abordagem proposta em Costa *et al.* (2019). oEm primeiro lugare os dados já são extraídos de portais de dados abertos ao invés de páginas Web. Em segundo lugare o objetivo do trabalho não está apenas na metodologia, mas na proposta de um repositório que irá conectar os dados acadêmicos de diversas instituições brasileiras, o DBAcademic. A seguir, os três passos apresentados na figura 3 serão detalhados.

3.1 EXTRAÇÃO DOS DADOS ABERTOS

Atendendo ao Decreto nº 8.777, de 11 maio de 2016, atualmente diversas instituições de ensino disponibilizam dados abertos. Essas instituições foram pesquisadasdno Portal Brasileiro de Dados Abertos, que indexa estes dados das instituições públicas.

Neste trabalho, identificou-se 45 instituições públicas de ensino com portais de dados abertos, sendo 25 universidades federais e 20 institutos federais. Entretanto, nessa primeira fase, foram incluídos os dados de 25 instituições que possuíam os conjuntos de dados mais acessíveis e completos. São elas: IFC, IFFAR, IFMA, IFMS, IFPA, IFPB, IFPI, IFRN, IFS, UFCA, UFCSPA, UFERSA, UFFS, UFMA, UFMS, UFOB, UFOP, UFPB, UFPEL, UFPI, UFRN, UFSJ, UFV, UNIFESSPA e UNIRIO.

Para a seleção dos dados, considerou-se os sete que eram os mais frequentes nos portais de dados abertos dessas instituições, como detalhado no quadro 1.

Quadro 1 − Descrições dos conjuntos de dados selecionados

Dados	Descrição
Docente	Informações de cada docente, como: nome, descrição, código, e-mail, áreas de interesses, departamento e URL para o currículo Lattes.
Curso	Informações de cada curso, como: nome, modalidade do curso, área de conhecimento do CNPQ, departamento, coordenador e título do profissional.
Departamento	Usualmente, os cursos e docentes são associados a um departamento. Desse modo, aqui incluem informações como: nome, localidade, chefe, centro no qual ele está associado e código.
Centro	Geralmente, o centro é uma unidade acadêmica em uma hierarquia superior aos departamentos. Os dados são, frequentemente: nome, localidade e diretor.
Grupo de Pesquisa	Os grupos de pesquisa são conjuntos de docentes e discentes que estudam um dado tema, tendo um docente como coordenador. Os dados são, usualmente: nome, área de conhecimento e coordenador.
Monografias (ou trabalhos de conclusão de curso)	Informações sobre as monografias (ou trabalhos de conclusão de curso) dos discentes, tais como: título, nome do aluno, nome do orientador, nome do curso, ano e data da defesa.
Discente	Este conjunto engloba as informações dos alunos ativos, ingressantes ou egressos da universidade. Geralmente, contém poucos atributos, como: nome, matrícula, período de ingresso e nome do curso.

Fonte: Elaborado pelo autor (2020).

Importante destacar aqui que, diferentemente de Costa *et al.* (2019), neste trabalho, não foram extraídos dados de páginas Web, mas sim de portais de dados abertos. Como foi discutido na seção 2, esses portais já disponibilizam dados em formatos abertos e acessíveis por algoritmos computacionais. Assim, para a extração, foi necessário apenas identificar os endereços desses recursos. Por exemplo, o recurso docente da Universidade Federal do Rio Grande do Norte é acessível através do seguinte endereço: http://dados.ufrn.br/api/action/datastore_search?resource_id=ff0a457e-76fa-4aca-ad99-48aebd7db070

Esses dados podem ser acessados por meio de navegadores Web, como o Chrome ou o Firefox. Contudo, para a automação, o ideal é escrever um código em alguma linguagem de programação que possa extrair e processar esses dados. Neste trabalho, foi utilizada a linguagem de programação Python[3].

3.2 TRANSFORMAÇÃO PARA DADOS CONECTADOS

Antes de descrever esse processo, é importante destacar que os dados conectados representam um paradigma diferente para a representação da informação. Nos portais de dados abertos, os dados são usualmente disponibilizados em formatos tabulares (linhas e colunas) ou como coleção de objetos com suas propriedades e valores. Os dados conectados são construídos a partir de três blocos básicos (ISOTANI; BITTENCOURT, 2015):

1. Modelo de dados padrão;
2. Vocabulários de referência;
3. Protocolo padrão de consulta.

Tanto o modelo de dados padrão quanto o vocabulário de referência utilizam, geralmente, o RDF que foi proposto e é mantido pela W3C para representação de metadados. Esse formato representa os dados como coleções de afirmativas (ou triplas) declaradas por um sujeito, um predicado e um objeto.

O sujeito e o objeto correspondem aos recursos a serem conectados, enquanto o predicado caracteriza a natureza dessa conexão direcionada do sujeito ao objeto. Um predicado também pode ser denominado de propriedade. Um objeto, em algum momento, pode ser um dado literal como um número, uma data ou um texto. Na seguinte tripla, por exemplo, o objeto é um texto, especificamente, o nome de Leonardo Da Vinci.

- **Sujeito:** http://dbpedia.org/page/Leonardo_da_Vinci
- **Predicado:** <http://dbpedia.org/ontology/birthName>
- **Objeto:** "Leonardo di ser Piero da Vinci"

Diferentemente do objeto, o sujeito precisa sempre ser um recurso e estar associado a um identificador único (URI). Neste trabalho, os sujeitos são recursos como docentes, discentes, cursos e departamentos. Sendo assim, é necessário definir um esquema para a criação de URIs para todos eles. Para isso, foi registrado um domínio onde cada recurso é associado ao seguinte esquema: www.dbacademic.tech/resource/<codigo único>. O código único foi gerado por meio do algoritmo de sintetização de mensagem MD5 (RIVEST, 1992). Em resumo, esse algoritmo retorna um código de 128 bits para um dado texto. Como entrada, usou-se um texto que é a concatenação (ou união) da sigla da instituição, o nome e um código do recurso. A sigla da universidade e o nome do recurso foram necessários para garantir um código único entre os recursos e as instituições.

Além dos sujeitos, os predicados também precisam estar associados a uma URI, que, nesse caso, representa um vocabulário. No exemplo anterior, o predicado birthName faz parte de um vocabulário criado pelo projeto DBpedia e está associado ao endereço: http://dbpedia.org/ontology/birthName. Muitos projetos precisam criar sua própria ontologia, que, no contexto computacional, são especificações formais e explícitas de conceitualizações compartilhadas e servem como base para garantir uma comunicação livre de ambiguidades (BREITMAN, 2006).

[3] Mais informações sobre essa linguagem podem ser encontradas em https://www.python.org/.

Além disso, segundo os quatro princípios dos dados conectados, eles devem, sempre que possível, se conectar a dados e vocabulários já existentes. Esse é um importante princípio, que permitiu construir uma base de dados global como a ilustrada pela figura 1. Desse modo, utilizou-se, como principal fonte de referência, as ontologias e os vocabulários citados nos estudos de Alencar *et al.* (2018), Costa *et al.* (2019); Kessler e Kauppinen, (2015), Piedra *et al.* (2014), Rocha e Lóscio (2015), Zablith, Fernandez e Rowe (2012) . Nesses trabalhos destacaram-se os vocabulários descritos no quadro 2:

Quadro 2 − Descrições dos vocabulários pesquisados

Vocabulário	Prefixo	Descrição
Academic Institution Internal Structure Ontology	AIISO	Descreve a estrutura organizacional interna de uma instituição acadêmica.
Dublin Core	DC	Descreve metadados genéricos.
Bibliographic Ontology Specification	BIBO	Descreve citações e referências bibliográficas.
Friend Of A Friend	FOAF	Descreve o vínculo de pessoas, seja por informações formais, como documentos físicos e digitais, seja por relacionamentos não formais.
Corresponde ao VCF (Virtual Contact File)	VCARD	Descreve pessoas e organizações utilizando *técnicas da web semântica.*
DBpedia Mappings Wiki	DBO	Descreve a semântica da extração dos dados da Wikipedia.
Open Information Center	OpenCIn	Descreve o ambiente acadêmico no qual o Centro de Informática da UFPE está imerso, essencialmente focalizado nos docentes e entidades relacionadas.
Web Ontology Language	OWL	Usado para representar e instanciar ontologias na World Wide Web.
Organization Ontology	ORG	Descreve estruturas organizacionais.

Fonte: Elaborado pelo autor (2020).

Para a seleção dos vocabulários, optou-se por aqueles mais consolidados e utilizados nos trabalhos anteriores. Contudo, como será descrito na seção 4, identificaram-se alguns desafios para o reuso de alguns deles. Dessa maneira, deverá ser proposta uma ontologia específica para as instituições de ensino em um trabalho futuro. Essa ontologia irá reusar muitos dos vocabulários apresentados no quadro 2, mas terá que incluir novos termos que representem melhor alguns conceitos das instituições de ensino. Além disso, espera-se que esse vocabulário possa ser construído por meio de parceria entre pesquisadores de diferentes instituições.

Após a definição de um esquema de geração de URIs e a seleção de vocabulários, é possível realizar a transformação entre os formatos de serialização. Estes são modelos de representação de dados que definem regras de sintaxe e esquemas para validação da informação. Os formatos de serialização mais comuns nos portais de dados abertos são: *JavaScript Object Notation* (JSON) e *Comma-Separated Values* (CSV). Em vista disso, foi necessária a transformação desses formatos em outro que fosse capaz de representar coleções de triplas, como o RDF/XML. Lembrando que o RDF é apenas um modelo de dados abstrato, enquanto o RDF/XML é um formato de serialização. A quadro 3 apresenta um fragmento de um documento RDF serializado como RDF/XML.

Quadro 3 − Representação de um docente por meio do RDF/XML

```
<rdf:Description rdf:about=”https://www.DBacademic.tech/resource/b39ba05094dd03dee6515349c07661a1”>

<foaf:name>LEVINDO DINIZ CARVALHO</foaf:name>

<owl:sameas rdf:resource=”https://sig.ufsj.edu.br/sigaa/public/docente/portal.jsf?siape=1943390”/>

<cin:siape>1943390</cin:siape>

<rdf:type rdf:resource=”http://dbpedia.org/ontology/Professor”/>

</rdf:Description>
```

Fonte: Elaborado pelo autor (2020).

A transformação foi realizada por intermédio da biblioteca *Simple Object-triple Mapping*[4] (SIMPOT), proposta em Costa *et al.* (2019). Uma vantagem dessa biblioteca é que ela permite fazer o mapeamento entre os modelos de objeto e tripla de forma declarativa. Essa abordagem tornou o processo de transformação mais fácil de replicar para os dados das diversas instituições selecionadas neste trabalho.

3.3 CARREGAMENTO E CONSULTA

Uma vez transformados em RDF/XML, foi necessário carregar os dados em um banco de dados específico para o armazenamento e a consulta de dados conectados, denominados, em inglês, como *triple-store*. Alguns exemplos incluem: Virtuoso, Apache Jena Fuseki, RDFFox e Neo4G. Em Rohloff *et al.* (2007), é apresentada uma análise de alguns desses sistemas. Neste trabalho, foi utilizado o Apache Jena Fuseki em razão de ele ter uma configuração mais simples e com suporte ao *Simple Protocol and RDF Query Language* (SPARQL). Esse protocolo permite que uma única consulta acesse várias bases de dados, tratando-as como se fossem um banco de dados global. Ele foi implantado na plataforma Heroku (www.heroku.com) e as consultas já podem ser realizadas no endereço http://sparql.dbacademic.tech/.

Na próxima seção, serão apresentados alguns resultados dessa primeira fase de desenvolvimento do repositório DBacademic, incluindo alguns exemplos de consultas.

4 – RESULTADOS

Foram extraídos sete conjuntos de dados de 25 instituições públicas de ensino, como detalhado na tabela 1. Observe que alguns conjuntos de dados estavam disponíveis em apenas algumas instituições.

Esses recursos foram então mapeados para conceitos, ou seja, classes de vocabulários já existentes. Os atributos também foram mapeados para propriedades e relacionados a termos de vocabulários, da mesma maneira, já existentes, como apresentados no quadro 4.

Como critério para a seleção dessas propriedades considerou-se a disponibilidade dessas informações nos diferentes portais. Mesmo assim, muitas delas não foram encontradas nos conjuntos de dados fornecidos pelas instituições, como o e-mail e o endereço para o currículo Lattes dos docentes.

Nessa fase, não foram desenvolvidos vocabulários específicos que fossem capazes de representar todos os dados das instituições de ensino. Focalizou-se apenas os dados extraídos e priorizou-se o reuso de vocabulários já existentes, entre eles, os apresentados no quadro 2 na seção 3.2. Contudo, alguns desses termos deverão ser revisados em trabalhos futuros, a exemplo do *Bibliographic Ontology Specification* (BIBO), uma ontologia bem consolidada para representar diversas publicações, incluindo dissertação de mestrado e tese de doutorado. Mesmo assim, nela não foi encontrado um conceito que pudesse representar com exatidão o que se denomina, no Brasil, trabalhos de conclusão de curso (ou monografias). Nessa fase de desenvolvimento, as monografias estão sendo mapeadas como bibo:Report, que não representa adequadamente o conceito de trabalho de conclusão de curso.

Na versão 0.1, a ontologia do DBacademic incluiu apenas os termos que foram reusados e o modo como eles se relacionam. Para melhor compreender as relações entre esses vocabulários, é possível consultar a versão mais atual no endereço http://purl.org/ontology/dbacademic.

Como exemplo, a figura 4 ilustra como uma monografia está relacionada a no mínimo outros quatro recursos.

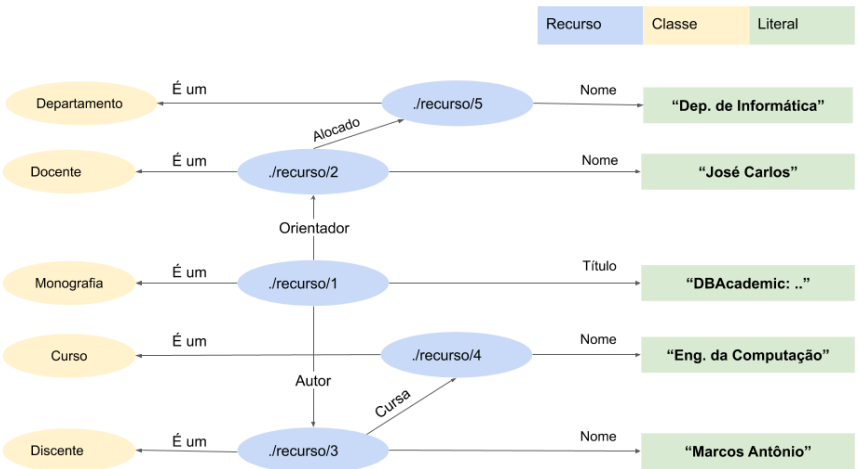
[4] Mais detalhes sobre essa biblioteca encontra-se em https://github.com/dbacademic/simpot

Tabela 1 – Recursos extraídos por instituição de ensino

Recursos	Instituições	Nº
Discente	UFMA, UFPI, UFRN, IFPA, IFMA, IFC, IFMS, IFRN, IFFAR	9
Docente	UFMA, UFPI, UFRN, UNIFESSPA, UFSJ, IFMA, IFPB, IFRN, IFMS, IFS, UFCSPA, UFV, UFMS, UFPEL, IFFAR	15
Curso	UFMA, UFPI, UFRN, UFPB, UFMS, IFMA, UFCA, UFCSPA, UFFS, UFPEL, IFMS, IFPB, IFRN, UFSJ, UFV, UNIFESSPA, UNIRIO, IFFAR	18
Centro	UFMA, UFPEL, UFRN,	3
Departamento	UFMA, UFRN, IFFAR	3
Grupo de pesquisa	UFRN, UFV, UFCA, IFC, UFPI, UFOP, UNIFESSPA, IFFAR, UFERSA	9
Monografia	UFMA, UFRN, UFOB	3
Total		60

Fonte: Elaboração do autor (2020).

Figura 4 – Representação gráfica da relação entre os recursos de uma monografia



Fonte: Elaboração do autor (2020).

Quadro 4 – Mapeamento de cada recursos para classes e propriedades de vocabulários existentes

Recurso	Classe	Propriedades
Docente	dbo:Professor	foaf:name, vcard:hasTelephone, vcard:hasPhoto, dbo:abstract, vcard:hasEmail, vcard:hasGender, owl:sameAs, cin:SIAPE, cin:academicDegree,cin:lattes, org:memberOf, owl:sameAs
Curso	aiiso:Programme	foaf:name, uai:hasKnowledgeArea, aiiso:responsibilityOf, aiiso:part_of, aiiso:code, owl:sameAs
Departamento	aiiso:Department	foaf:name, aiiso:responsibilityOf, owl:sameAs, aiiso:code, aiiso:part_of, owl:sameAs
Centro	aiiso:Center	foaf:name, aiiso:responsibilityOf,owl:sameAs, aiiso:code, owl:sameAs
Grupo de Pesquisa	aiiso:ResearchGroup	foaf:name, aiiso:hasKnowledgeArea, aiiso:responsibilityOf, owl:sameAs
Discente	cin:Student	foaf:name, dc:identifier, cin:isMemberOf, owl:sameAs
Monografia	bibo:Report	dc:title, dc:creator, bibo:issuer, dc:contributor, dc:date, owl:sameAs

Fonte: Elaboração do autor (2020).

Como ilustra a figura 4, o */recurso/1* é uma monografia que está relacionada a um docente *(./recurso/2)* e a um discente *(./recurso/3)*. Uma característica importante dos dados conectados é a possibilidade de um recurso ser facilmente conectado a recursos de outras bases de dados. O *./recurso/4,* por exemplo, poderia estar associado ao recurso http://pt.dbpedia.org/resource/Universidade_Federal_do_Maranhão, que pertence à base de dados do DBpedia.

Depois de extraídos e transformados, esses dados foram importados para um banco de dados conectado, resultando em 884.838 triplas. A tabela 2 apresenta a quantidade de triplas por cada classe.

Tabela 2 – Quantidade de triplas associadas à cada classe

Classe	Nº Triplas
cin:Student	132.936
bibo:Report	61.875
dbo:Professor	24.564
aiiso:ResearchGroup	2.917
aiiso:Programme	2.367
aiiso:Department	956
aiiso:Center	53

Fonte: Elaboração do autor (2020).

As instituições que tiveram mais dados carregados foram: UFRN, IFRN e UFMA. A tabela 3 apresenta as dez instituições que possuem mais dados.

Tabela 3 – Número de triplas por instituição

Instituição	Nº Triplas
UFRN	71415
IFRN	41173
UFMA	40133
IFMA	23565
UFPA	15950
IFFAR	10019
UFV	4178
UFPI	3751
UFSJ	3445
UFMS	2234

Fonte: Elaboração do autor (2020).

Os dados conectados são consultados por meio de um SPARQL *endpoint*, que é provido pelo servidor de dados conectados, ou *triple store*. Para este trabalho, os dados foram carregados e implantados no servidor Apache Jena Fuseki e estão disponíveis no endereço http://sparql.dbacademic.tech/.

Através deste endereço, é possível escrever e enviar as consultas diretamente, usando um navegador Web. Como exemplo, o quadro 5 apresenta um código que irá retornar ordenadamente a quantidade de cursos de engenharia por instituição de ensino. Essa consulta pode ser acessada também no endereço encurtado https://bit.ly/35lAA7d.

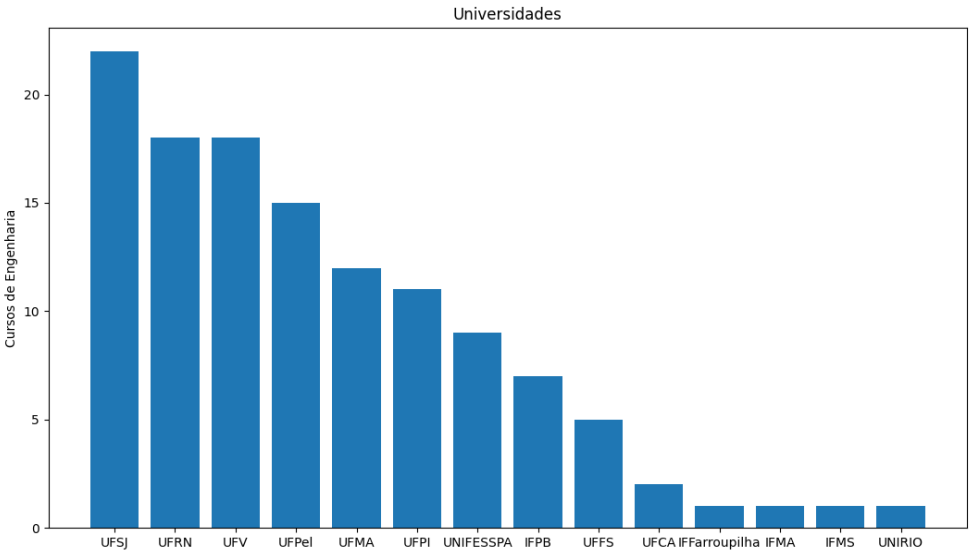
Quadro 5 – Consulta SPARQL que retorna a quantidade de cursos de engenharia por instituição

```
PREFIX dbo: <http://dbpedia.org/ontology/>

PREFIX aiiso: <http://purl.org/vocab/aiiso/schema#>

PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?sigla (COUNT(?curso) as ?quant_cursos)
WHERE {

?curso a aiiso:Programme.

?curso aiiso:part_of ?instituicao.

?instituicao foaf:nick ?sigla.

?instituicao a dbo:EducationalInstitution.

FILTER regex(lcase(?curso_name), "engenharia") }

GROUP BY ?sigla

ORDER BY DESC (?quant_cursos)
```

Fonte: Elaboração do autor (2020).

Além da possibilidade de consultar os dados diretamente pelo navegador Web, é possível enviar as consultas por intermédio de um aplicativo móvel ou de um algoritmo de análise e visualização de dados. A título de exemplo, um código escrito na linguagem Python pode enviar a consulta do quadro 5 e gerar um gráfico a partir do resultado. Esse código pode ser acessado e executado diretamente pelo Google Colab (https://colab.research.google.com/) no endereço https://bit.ly/2y2aTfZ. Ao executar esse código é gerada a saída apresentada na figura 5.

Figura 5 – Gráfico produzido de uma consulta SPARQL ao DBAcademic



Fonte: Elaboração do autor (2020).

A figura 5 apresenta um gráfico de barras com a quantidade de cursos de engenharia por instituição de ensino, apontando a UFSJ com a maior quantidade de cursos de Engenharia, lembrando que esse resultado é gerado a partir dos dados que estão no DBacademic e que não inclui ainda todas as instituições de ensino.

Uma das vantagens dos dados conectados, discutidos na seção 2, é a possibilidade de incluir os dados de outras bases de dados, como o DBpedia. No exemplo do quadro 6, uma instituição na base de dado do DBacademic é associada a um recurso na DBpedia. Com essa associação, é possível chegar a propriedade dbo:city, que levu a outro recurso, que permitu acessar o nome dessa cidade. Esse nome é então inclnído nos resultados dessa consulta, que é realizada a partir do SPARQL *endpoint* do DBacademic.

Quadro 6 – Consulta SPARQL que permite a integração de bases de dados conectados

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT distinct ?instituicao ?nome_cidade
WHERE {
  ?instituicao a dbo:EducationalInstitution.
  ?instituicao owl:sameAs ?same.
  SERVICE <http://dbpedia.org/sparql> {
    ?same dbo:city ?cidade.
    ?cidade foaf:name ?nome_cidade    }
  FILTER REGEX(str(?same), "^http://dbpedia.org")}
```

Fonte: Elaboração do autor (2020).

Do mesmo modo que, no quadro 6, foi incluído o DBpedia, seria possível incluir diversas outras bases de dados, enriquecendo as análises e a visualização dos dados.

Por meio desses resultados, é possível identificar o potencial do DBacademic de integração de dados, valorizando ainda mais os dados que já estão abertos pelas diversas instituições de ensino. Contudo, para a sua ampliação, deverão ser considerados os três grandes desafios a seguir.

A NECESSIDADE DE ELABORAÇÃO DE UMA ONTOLOGIA MAIS ADEQUADA

A criação de ontologias é sempre um grande desafio, e poderá ser muito beneficiada com a participação de outras instituições de ensino. No Brasil, foram identificados dois projetos que avançaram na proposta de uma ontologia que faz o reuso de diversos vocabulários (ALENCAR *et al.*, 2018; ROCHA; LÓSCIO, 2015). Desse modo, para o próximo passo, será necessário compatibilizar e adequar melhor as propostas existentes para atender às demandas do Dbacademic.

A MELHORIA NA QUANTIDADE E QUALIDADE DOS DADOS

Mesmo sem realizar uma análise quantitativa e qualitativa dos portais de dados abertos, foi possível perceber a ausência e a falta de atualização em muitos deles. Em geral, o que está disponível nos portais são apenas visões dos dados, definidas pelo detentor deles. No paradigma de dados conectados, armazenam-se apenas os dados e suas conexões. As visões são criadas por meio de consultas, que poderiam incluir dados de outras bases, como destacado na seção 4.

A NECESSIDADE DE UM PLANO PARA A MANUTENÇÃO E INSTITUCIONALIZAÇÃO DO PROJETO

Na literatura, existem diversas iniciativas similares e relevantes que atualmente estão tendo dificuldades de serem mantidas. Muitos portais de dados conectados já não estão mais em funcionamento e suas ontologias não estão mais disponíveis. Essa dificuldade pode ser ainda maior, devido à necessidade de manter atualizados os dados de diversas instituições.

5 – CONCLUSÕES

Este artigo apresentou um projeto que tem o objetivo de conectar dados abertos de diversas instituições em um grande repositório de dados, denominado DBacademic. Os resultados dessa primeira fase do projeto mostraram o grande potencial dessa proposta, que poderá se tornar uma importante referência, principalmente, para consultas a dados entre instituições de ensino.

Como resultados dessa primeira fase, o DBacademic já conseguiu incluir sete conjuntos de dados de 25 instituições de ensino do Brasil. Atualmente, já é possível realizar diversas consultas em sua base de dados com quase 900 mil triplas. Não foi realizada ainda qualquer avaliação de eficiência dessas consultas, visto que o servidor atualmente implantado utiliza uma quota gratuita com algumas limitações de eficiência. Espera-se que, ao finalizar essa fase de teste, consigamos implantar o repositório em um servidor dedicado e mantido por uma das instituições de ensino.

Além dos resultados, deve-se destacar os três principais desafios: necessidade de elaboração de uma ontologia mais adequada; melhoria na qualidade dos dados; e necessidade de um plano para a manutenção e institucionalização do projeto. Todos esses desafios poderão ser enfrentados por meio de parcerias com instituições e pesquisadores. Nesse sentido, um dos objetivos subjacentes deste artigo é apresentar e validar esse projeto com a comunidade acadêmica para fazer avançar as parcerias.

Além dos três desafios, um estudo relevante a ser considerado em um trabalho futuro é a análise da efetividade dos Planos de Dados Abertos elaborados pelas instituições, incluindo a identificação de quais desafios elas estão enfrentando para seguir os oito princípios fundamentais dos dados abertos propostos pelo open Government Working Group (2007).

REFERÊNCIAS

ALENCAR, A.; XAVIER, D.; CHAVES, L. C.; SOUZA, D. Y. Publicação e consumo de dados abertos conectados acadêmicos. *Revista Principia:* Divulgação Científica e Tecnológica do IFPB, v. 1, n. 42, p. 136, 18 ago. 2018. DOI 10.18265/1517-03062015v1n42p136-145. Disponível em: http://periodicos.ifpb.edu.br/index.php/principia/article/view/1988. Acesso em: 11 mar. 2021.

AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. DBpedia: A Nucleus for a Web of Open Data. *In*: ABERER, K.; CHOI, K.-S.; NOY, N.; ALLEMANG, D.; LEE, K.-I.; NIXON, L.; GOLBECK, J.; MIKA, P.; MAYNARD, D.; MIZOGUCHI, R.; SCHREIBER, G.; CUDRÉ-MAUROUX, P. (orgs.). *The Semantic Web*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4825, p. 722–735. DOI 10.1007/978-3-540-76298-0_52. Disponível em: http://link.springer.com/10.1007/978-3-540-76298-0_52. Acesso em: 11 mar. 2021.

BERNERS-LEE, T. *Linked Data*. 2009. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 13 jan. 2021.

BERTIN, P. R. B.; MACHADO, C. D.; VISOLI, M. C.; DRUCKER, D. P.; PINTO, D. M. A construção do Plano de Dados Abertos de uma organização pública de Pesquisa e Desenvolvimento e o desafio de uma Ciência Agropecuária Aberta. *Revista Eletrônica de Comunicação, Informação e Inovação em Saúde*, v. 11, 30 nov. 2017. DOI 10.29397/reciis.v11i0.1411. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1411. Acesso em: 11 mar. 2021.

BRASIL. *Decreto nº 8.777, de 11 de maio de 2016.* Institui a Política de Dados Abertos do Poder Executivo federal. 11 maio 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm. Acesso em: 22 fev. 2021.

BRASIL. *Lei nº 12.527, de 18 de novembro de 2011 [Lei de Acesso à Informação]*. 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. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 30 mar. 2020.

BRASIL. *Portal Brasileiro de Dados Abertos*. 2019. Disponível em: http://dados.gov.br. Acesso em: 13 set. 2019.

BRASIL. MINISTÉRIO DA DEFESA. Sobre a Lei de Acesso à Informação. 9 set. 2020. *Gov.br*. Disponível em: https://www.gov.br/defesa/pt-br/acesso-a-informacao/outros/sobre-lei-de-acesso-a-informacao. Acesso em: 11 mar. 2021.

BREITMAN, K. K. *Web semântica: a internet do futuro*. Rio de Janeiro: LTC, 2006.

CAROSSI, D. F.; TEIXEIRA FILHO, J. G. de A. Uma Análise dos Pedidos de Acesso à Informação Encaminhados a uma Instituição de Ensino Superior. *Gestão.Org*, v. 14, n. 2special, p. 255–264, 1 maio 2017. DOI 10.21714/1679-18272016v14Esp2.p255-264. Disponível em: http://www.revista.ufpe.br/gestaoorg/index.php/gestao/article/viewFile/906/528. Acesso em: 11 mar. 2021.

COSTA, I. N. da; ANDRADE, L. do E. S.; RESENDE, L.; TONIN, P.; COSTA, M.; SANTOS, Z. *Manual da Lei de Acesso à Informação para Estados e Municípios*. Brasília: CGU, 2013(Brasil transparente). Disponível em: https://acessoainformacao.valparaisodegoias.go.gov.br/res/docs/manual_lai_estadosmunicipios.pdf. Acesso em: 11 mar. 2021.

COSTA, S. S.; SOUSA, M. V. D.; SILVA, M. L. da; OLIVEIRA, E. C. de; GUIMARÃES, J. V. M. Uma solução semi-automática para extração, transformação e carga de dados abertos conectados. *In*: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 2019. *Anais* [...]. Brasília: FCI, 2019. p. 138–143. Disponível em: http://widat2019.fci.unb.br/index.php/anais-widat-2019. Acesso em: 11 mar. 2021.

GAMA, J. R.; RODRIGUES, G. M. Transparência e acesso à informação: um estudo da demanda por informações contábeis nas universidades federais brasileiras. *Transinformação*, v. 28, n. 1, p. 47–58, abr. 2016. DOI 10.1590/2318-08892016002800004. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862016000100047&lng=pt&tlng=pt. Acesso em: 11 mar. 2021.

HEATH, T.; BIZER, C. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, v. 1, n. 1, p. 1–136, 9 fev. 2011. DOI 10.2200/S00334ED1V01Y201102WBE001. Disponível em: http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001. Acesso em: 11 mar. 2021.

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

KESSLER, C.; KAUPPINEN, T. Linked Open Data University of Münster – Infrastructure and Applications. *In*: SIMPERL, E.; NORTON, B.; MLADENIC, D.; DELLA VALLE, E.; FUNDULAKI, I.; PASSANT, A.; TRONCY, R. (orgs.). *The Semantic Web: ESWC 2012 Satellite Events*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015. v. 7540, p. 447–451. DOI 10.1007/978-3-662-46641-4_43. Disponível em: http://link.springer.com/10.1007/978-3-662-46641-4_43. Acesso em: 11 mar. 2021.

MCCRAE, J. P.; ABELE, A.; BUITELAAR, P.; CYGANIAK, R.; JENTZSCH, A.; ANDRYUSHECHKIN, V.; DEBATTISTA, J.; NASIR, J. *The Linked Open Data Cloud.* 20 maio 2020. Disponível em: https://lod-cloud.net/. Acesso em: 11 mar. 2021.

OPEN GOVERNMENT WORKING GROUP. *The 8 Principles of Open Government Data*. 2007. Disponível em: https://public.resource.org/8_principles.html. Acesso em: 11 mar. 2021.

OPEN KNOWLEDGE FOUNDATION. *Definição de Conhecimento Aberto*. 2019. Disponível em: https://opendefinition.org/od/2.0/pt-br/. Acesso em: 9 set. 2020.

PIEDRA, N.; TOVAR, E.; COLOMO-PALACIOS, R.; LOPEZ-VARGAS, J.; ALEXANDRA CHICAIZA, J. Consuming and producing linked open data: the case of OpenCourseWare. *Program*, v. 48, n. 1, p. 16–40, 28 jan. 2014. DOI 10.1108/PROG-07-2012-0045. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PROG-07-2012-0045/full/html. Acesso em: 11 mar. 2021.

RIVEST, R. *The MD5 Message-Digest Algorithm*, n. RFC1321. [*S. l.*]: RFC Editor, abr. 1992. DOI 10.17487/rfc1321. Disponível em: https://www.rfc-editor.org/info/rfc1321. Acesso em: 11 mar. 2021.

ROCHA, J.; LÓSCIO, B. OpenCIn: Usando Dados Abertos e Conectados para a Publicação de dados sobre o CIn/UFPE. *In*: CONCURSO DE TRABALHOS DE INICIAÇÃO CIENTÍFICA DA SBC (CTIC-SBC), 34., 2015. *Anais* [...]. Recife: Sociedade Brasileira de Computação, 2015. v. 34, p. 11–20. Disponível em: https://sol.sbc.org.br/index.php/ctic/article/view/10014. Acesso em: 11 mar. 2021.

ROHLOFF, K.; DEAN, M.; EMMONS, I.; RYDER, D.; SUMNER, J. An Evaluation of Triple-Store Technologies for Large Data Stores. *In*: MEERSMAN, R.; TARI, Z.; HERRERO, P. (org.). *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. v. 4806, p. 1105–1114. DOI 10.1007/978-3-540-76890-6_38. Disponível em: http://link.springer.com/10.1007/978-3-540-76890-6_38. Acesso em: 11 mar. 2021.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. *Ciência da Informação*, v. 48, n. 3 (Supl.), p. 38–46, dez. 2019. Disponível em: http://revista.ibict.br/ciinf/article/view/4866/4428. Acesso em: 11 mar. 2021.

VASSILIADIS, P.; SIMITSIS, A.; SKIADOPOULOS, S. Conceptual modeling for ETL processes. *In*: THE 5TH ACM INTERNATIONAL WORKSHOP, 2002. *Proceedings* [...]. McLean, Virginia, USA: ACM Press, 2002. p. 14–21. DOI 10.1145/583890.583893. Disponível em: http://portal.acm.org/citation.cfm?doid=583890.583893. Acesso em: 11 mar. 2021.

ZABLITH, F.; FERNANDEZ, M.; ROWE, M. The OU Linked Open Data: Production and Consumption. *In*: GARCÍA-CASTRO, R.; FENSEL, D.; ANTONIOU, G. (orgs.). *The Semantic Web: ESWC 2011 Workshops*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 7117, p. 35–49. DOI 10.1007/978-3-642-25953-1_4. Disponível em: http://link.springer.com/10.1007/978-3-642-25953-1_4. Acesso em: 11 mar. 2021.

ZORZAL, L.; RODRIGUES, G. M. Transparência das informações das universidades federais: estudo dos relatórios de gestão à luz dos princípios de governança. *Biblios:* Journal of Librarianship and Information Science, n. 61, p. 1–18, 14 mar. 2016. DOI 10.5195/BIBLIOS.2015.253. Disponível em: http://biblios.pitt.edu/ojs/index.php/biblios/article/view/253. Acesso em: 11 mar. 2021.

AGRADECIMENTOS

Agradecemos aos autores dos projetos OpenCIN e OpenUAI pela comunicação, que foi breve, mas essencial a este trabalho. Espera-se que este trabalho seja apenas um ponto inicial para futuras parcerias