Migrando dos dados abertos para dados conectados: uma proposta para a Universidade Federal do Maranhão

Eddye Cândido de Oliveira[1], **José Victor Meireles Guimarães**[1], **Sérgio Souza Costa**[1]

[1] Curso de Engenharia da Computação – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brazil

eddyeoliver@gmail.com, jvictormguimaraes@gmail.com, sergio.costa@ufma.br

Resumo. *Este artigo descreve uma metodologia para (1) a extração dos dados abertos e públicos da Universidade Federal do Maranhão, (2) o seu mapeamento para dados conectados e (3) a sua publicação em um servidor de dados conectados. Para os experimentos foram utilizados quatro recursos: docente, discente, monografia e curso. Estes foram representados através de vocabulários bem aceitos e estabelecidos pela comunidade. Como resultado, estes dados foram carregados para um servidor da AWS (Amazon Web Server) e estão disponíveis para consulta no formato SPARQL, facilitando buscas complexas, conectando os diversos recursos, sem a necessidade de requisições múltiplas, como ocorre no caso de APIs para dados abertos.*

Abstract. *This paper describes a methodology for (1) the extraction of public open data of the Federal University of Maranhão, (2) their mapping to Linked open data and (3) their publication on a Linked Data server. For the experiments, four resources were used: teacher, student, monograph and major. These were represented by vocabularies well accepted and established by the community. As a result, this data was loaded to an AWS server (textit Amazon Web Server) and is available for viewing in SPARQL format, making it easier for complex searches, connecting the various resources, without the need for multiple requests, such as in APIs for open data.*

1. Introdução

De acordo com [Rowley 2007] os dados são definidos como símbolos que representam as propriedades dos objetos, eventos e seu ambiente. São produtos da observação, seja de um sensor eletrônico ou mesmo provenientes de pessoas compartilhando suas informações. Por exemplo, o aplicativo Google Maps informa a quantidade de tráfego nas vias com base nas informações coletadas dos usuários do seu sistema. Neste cenário, as pessoas passam a ser como sensores alimentando grandes sistemas. Esse grande volume de dados trouxe novas questões e problemas científicos, que demandou uma área específica denominada de "ciência dos dados". Além disso ela está impondo uma nova forma de fazer ciências, que foi designado como o quarto paradigma por Tony Hey [Hey et al. 2009]. No Brasil, as universidades são instituições com um grande volume de dados disponíveis através dos seus sites públicos. Contudo, muitas destas informações não são legíveis a algoritmos, tornando difícil seu acesso.

Uma das questões de interesse das ciências dos dados é como conectar dados de diferentes fontes gerando uma grande base de dados global. Essa questão foi colocada

pelo mesmo criador da web, Tim Berners-Lee [Berners-Lee 2006]. Nesse artigo, o autor apresentou uma nova web, que, diferentemente da anterior, iria conectar os dados em vez das páginas. Enquanto a primeira foi projetada para se utilizada por pessoas, a web dos dados estaria preparada para ser utilizada por algoritmos. Esses algoritmos poderiam extrair informações automaticamente de distintas bases de dados. Tim Berners-Lee propôs ainda uma classificação para os dados, onde a maior nota era dada para aquele dado que estivesse conectado com outras bases já existentes. A *dbpedia* é atualmente a base de dados mais conectada, e está funcionando como um elo de ligação entre diversas bases de dados do mundo todo. Por exemplo, se todas as bases de dados das universidades se conectarem a *dbpedia* para indicar sua cidade sede. Em um dado momento, seria possível fazer consultas questionando quais são as universidades e—ou cursos de uma dada cidade.

Considerado o potencial da integração dos dados através deste paradigma. O objetivo deste trabalho é ser a base de um projeto para a criação de uma grande base de dados conectados para a Universidade Federal do Maranhão.

Este artigo está organizado da seguinte maneira: a Seção 2 apresenta brevemente alguns conceitos sobre dados abertos e conectados, linguagem SPARQL; a Seção 3 apresenta a metodologia desenvolvida, a Seção 4 apresenta alguns resultados alcançados, esta última apresentando as conclusões.

2. Fundamentação

2.1. Dados abertos e dados conectados

Para entender a diferença entre dados abertos e dados abertos conectados (LOD) é preciso perceber a métrica utilizada para classificar dados na WEB. Segundo Tim Berners-Lee [Berners-Lee 2006], existem quatro regras para a publicação de dados, sua classificação dependerá do atendimento dessas regras, que é feita por meio de um sistema que concede uma nota que vai de uma a cinco estrelas. Para um determinado conjunto de dados ser considerado aberto é requerido ter pelo menos quatro estrelas na classificação, ou seja, atender três das quatro regras, e cinco estrelas para ser considerado dado aberto conectado, ou seja, atender todas as regras. Dito isso, têm-se as regras a seguir:

1. Usar URIs para nomeação;
2. Usar HTTP URI para permitir indexação;
3. Utilizar os padrões (RDF e SPARQL) para prover informações úteis;
4. Incluir links para outras URIs, encadeando os dados.

2.2. Dados abertos conectados

O formato RDF, segundo a W3C é um formato de dados para a representação da informação na WEB, armazena os dados em formato de triplas cujas quais possuem sujeito, predicado e objeto. Onde o sujeito sempre é uma URI e o objeto pode ou não ser uma *string* literal; enquanto o predicado é uma URI e descreve a relação entre o objeto e o sujeito. Fazendo isso, cada unidade de dado RDF ficará posicionado no grafo e terá relações com outros dados, como nós encadeados, conectados entre si por mais nós. Este modelo de publicação de dados tem como grande vantagem a sua potencialização pela colaboração coletiva e descentralizada. Uma vez que um determinado conjunto de

dados está publicado desta forma, seu conteúdo passa a fazer parte da WEB semântica, constituída por todos os dados conectados na internet. Tendo isso em mente, a interdependência desses dados os levam a ser solo fértil para a aplicação de métodos que visam a inferir informações ocultas na superfície, possibilitando consultas ágeis e eficientes. Na prática a WEB semântica funciona como um grande banco de dados global.

2.3. Serviço de Consulta

Na Web Semântica, os dados são representados através do modelo RDF, como já mencionado, e que estão armazenados em um banco de dados de triplas (*triple stories*). Assim como o banco de dados relacional tem sua pesquisa com o uso da linguagem SQL (*Structured Query Language*), o SPARQL realiza a consulta de dados representados em RDF na Web 3.0.

De acordo com Tim Berners-Lee [Berners-Lee 2006] "Tentar usar a Web Semântica sem SPARQL é como tentar usar um banco de dados relacional sem SQL". Definida padrão oficialmente pelo W3C em 2008, SPARQL quebra os padrões para Linguagens de consulta estabelecidas até então, tornando-se ferramenta indispensável para a utilização da Web Semântica.

Segundo Dave [Becket 2011], um de seus criadores, a sigla significa "*Simple Protocol and RDF Query Language*"(Linguagem e protocolo RDF simples). Em termos de funcionalidade, SPARQL é uma linguagem de consulta de dados para a Web Semântica, criada para realizar consultas e manipular dados estruturados no formato RDF; entretanto pela própria natureza descentralizada da Web Semântica, a comparação da linguagem com maneiras mais tradicionais de consulta, como dados relacionais, apesar de natural, é imprecisa.

A Web Semântica tem como objetivo permitir o compartilhamento, a combinação e a reutilização de dados em escala global. SPARQL não é limitada, é capaz de numa única consulta acessar várias bases de dados, com custo baixo, e por não estar atrelada a nenhum formato de dados específico, fornece ricos resultados.

3. Metodologia

A UFMA atualmente já possui um repositório de dados abertos, e em [Silva et al. 2017] os autores extraíram alguns dados públicos e disponibilizaram um serviço REST para serem acessados. Contudo, o acesso a essas informações demanda escrever códigos em linguagens de programação, além de exigir diversas requisições a cada recurso individualmente, consumindo recursos de rede. Esse trabalho tem como objetivo expor os dados da Universidade Federal do Maranhão como uma coleção de dados conectados. A Figura 1 apresenta então a metodologia desenvolvida neste trabalho dividida em quatro etapas: 1) modelagem, 2) extração e geração de dados, 3) armazenamento e 4) acesso. Estas etapas serão brevemente descritas a seguir.

3.1. Modelagem

A primeira etapa do trabalho consiste na modelagem das entidades que serão mapeadas como recursos RDFs. Neste trabalho, foram selecionadas apenas as entidades curso, discente, docente, publicação, unidade e subunidade. Quando se trata da criação de um conjunto de dados em Linked Data é importante associar vocabulários aos seus

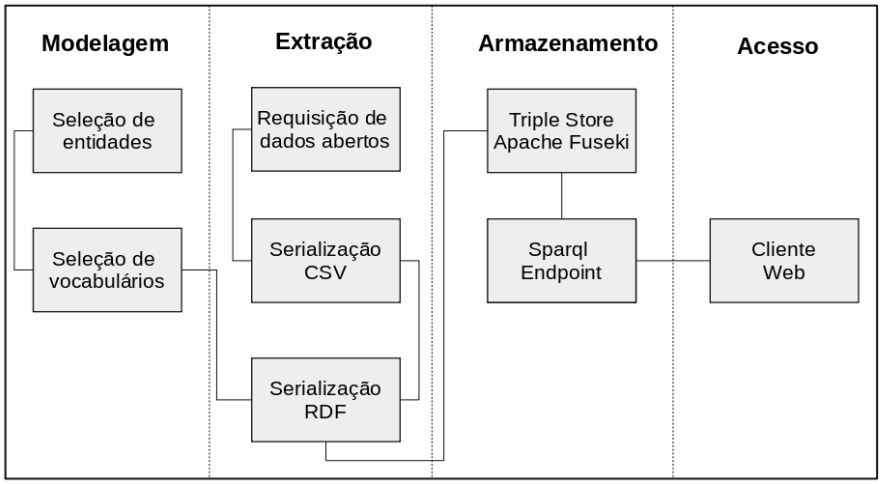

Figura 1. Visão ampla sobre a metodologia

dados ou entidades. Vocabulários são descrições de conceitos de um domínio (geral ou específico), materializadas em um conjunto de termos. Os termos de um vocabulário, juntamente com os seus relacionamentos, podem ser descritos por meio de ontologias e seus componentes (classes e propriedades), e este pode pode ser reusado através de URI [Batista and Lóscio 2013]. Neste trabalho foram reusados vocabulários consolidados na literatura, como FOAF, DUBLIN CORE e AIISO. Por exemplo, o vocabulário *Friend of a Friend*[1] foi usado para os atributos das entidades docente e discente. Porém, para a classificação das entidades discente, docente e curso, utilizou-se o vocabulário *Teaching Core Vocabulary Specification*. A Figura 2 apresenta as principais ligações entre as entidades, não destacando outros atributos literais, por exemplo o foaf:name que é um atributo do docente e discente. Os detalhes sobre os vocabulários, incluindo atributos e classes, serão disponibilizados em uma página sobre o projeto (`https://inovacampus.github.io/ludufma/`).

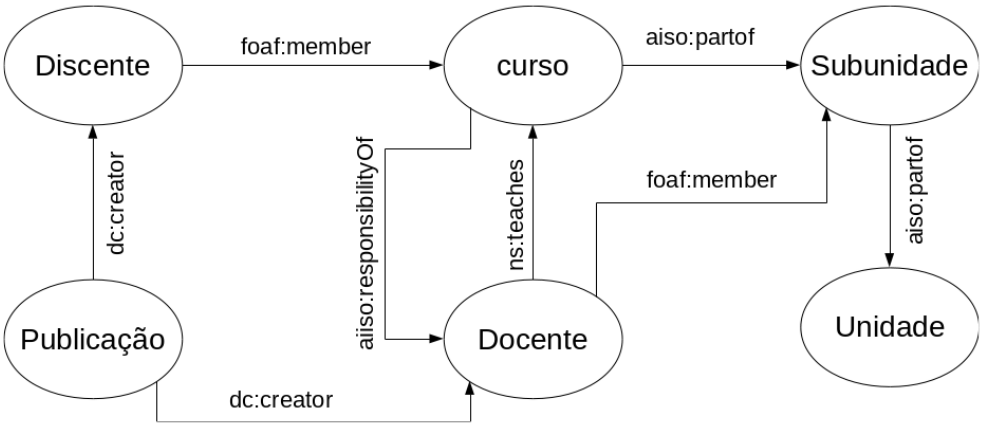

Figura 2. Entidades e principais conexões

[1]FOAF:-url aqui

Definidas as entidades, atributos e vocabulários, a próxima etapa consiste na extração dos dados.

3.2. Extração e geração dos dados

Para este trabalho foram utilizadas duas fontes de dados. A primeira é o site oficial de dados abertos da UFMA[2]. O segundo é um site não oficial desenvolvido em [Silva et al. 2017] e disponibilizado em https://dados-abertos-ufma.herokuapp.com/. Ambos sites ainda possuem poucos dados, porém foram suficientes para essa prova de conceitos.

O primeiro passo foi a extração dos dados para o formato CSV, através de alguns códigos escritos na linguagem Python (https://github.com/inovacampus/ludufma/tree/master/codigos). Estes dados serviram então de entrada do software Open Refine [3]. Esse software possui diversos recursos para lidar com dados tabulares, inclusive uma extensão específica para a geração de dados conectados. Com este software foi possível exportar os dados no formato RDF, como o exemplo da Figura 3.

```
<rdf:Description rdf:about="http://lud.ufma.br/person/siape1111">
      <rdf:type rdf:resource=
         "http://linkedscience.org/teach/ns#Teacher"/>
      <foaf:name>ARTUR ANTONIO DA ROCHA</foaf:name>
      <foaf:homepage rdf:resource="http://lattes.cnpq.br/1573092"/>
      <foaf:interest>Direito Empresarial</foaf:interest>
      <foaf:member rdf:resource=
      "http://lud.ufma.br/organizational/01_1010"/>
</rdf:Description>
```

Figura 3. Exemplo de um dado RDF gerado para um discente

Os dados nos formatos CSV e RDF, podem ser acessados no seguinte endereço: https://github.com/inovacampus/ludufma/tree/master/dados.

3.3. Armazenamento

Após a serialização dos dados no formato RDF, estes foram enviados para um servidor triple store. Para este trabalho foi utilizado o servidor Apache Jena Fuseki [4]. Ele foi escolhido para este trabalho por ser um servidor simples e de fácil instalação e configuração. Este servidor está rodando em um servidor da Amazon no seguinte endereço http://ec2-3-16-56-184.us-east-2.compute.amazonaws.com:3030/d:3030/.

4. Resultados

A linguagem SPARQL permite realizar buscas complexas com apenas uma única requisição, diferentemente de uma arquitetura orientada a recurso como o REST. Nessa

[2]Dados abertos UFMA:http://dadosabertos.ufma.br/
[3]Maisinformaμesemhttp://openrefine.org/
[4]Mais informações em https://jena.apache.org/documentation/fuseki2/

arquitetura cada entidade representa um recurso e tem o seu *endpoint* próprio. Considere uma consulta base de dados de monografia, que retorne o título, nome do discente, nome do orientador, currículo Lattes do orientador e o nome do curso. Contudo, apenas dos orientadores que tem como área de interesse ”História”. Traduzir essa consulta para um código em uma linguagem de programação demanda requisições a diferentes recursos, além de produzir um código difícil de entender e de adaptar. Por outro lado, a mesma consulta poderia ser feita com a seguinte declaração em SPARQL da Figura 4.

```
SELECT ?titulo ?discente ?orientador ?lattes ?curso
WHERE {
    ?m dc:title ?titulo.
    ?m dc:creator ?url_discente.
    ?m dc:contributor ?url_docente.
    ?url_discente foaf:name ?discente.
    ?url_discente foaf:member ?url_curso.
    ?url_curso ns:courseTitle ?curso.
    ?url_docente foaf:name ?orientador.
    ?url_docente foaf:interest ?interesse_orientador.
    ?url_docente foaf:homepage ?lattes.
    FILTER regex(?interesse_orientador, "Historia")
}
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 6 of 6 entries

Search: Show 1000 entries

	titulo	discente	orientador	interesse_orientador	curso
1	"QUILOMBO URBANO: POLÍTICA, CULTURA E RESISTÊNCIA NO HIP HOP DO MARANHÃO"	"WHERLYSHE SOUSA DE MORAIS"	"ANTONIO EVALDO ALMEIDA BARROS"	"História Social da Cultura. Estado e Cultura. Estudos Étnicos. História da África (África do Sul e Moçambique). Festa, Memória, Patrimônio, Educação e Cidadania. Educação do Campo. Educação para as Relações Étnico-Raciais. História Social das Ideias. Filosofia e pensamento africanos. Produção da (in)diferença, Liberdade e Cidadania no mundo contemporâneo."	"CURSO DE CIÊNCIAS HUMANAS - SOCIOLOGIA / CAMPUS III"

Figura 4. Código e resultado de um consulta SPARQL

Um outro experimento realizado foi o desenvolvimento de uma página html que realiza uma requisição enviando uma consulta sparql para um servidor. A página pode ser testada através da seguinte url `http://ec2-3-16-56-184.us-east-2.compute.amazonaws.com/monografias.html` e seu código fonte está disponível em: `https://github.com/inovacampus/ludufma/blob/master/codigos/monografias.html`.

Observou-se, durante a extração de dados, que poucos docentes possuem informações cadastradas em seu perfil. Isso se deve ao fato de que os dados utilizados são resultados de um *scrapping* realizados nas páginas públicas, e muitas delas são responsabilidades dos professores, que ainda não estão aproveitando adequadamente este recurso.

5. Conclusão

Tendo em mente as vantagens apresentadas por este trabalho ao transformar os dados abertos em dados conectados. Mesmo com um pequeno volume de dados, já foi possível verificar o potencial desta arquitetura. Uma grande vantagem dos dados conectados é a facilidade de integração com distintas bases e o uso da linguagem SPARQL para consulta de dados. Porém, para expandir o volume de dados serão necessários os seguintes passos:

- Identificar outras entidades e atributos nas páginas públicas da UFMA.
- Evoluir a API de dados abertos para realizar o *scrapping* destes novos dados.
- Desenvolver um sistema que acesse as APIs de dados abertos, gerem os dados em RDF e enviem a um servidor *Triple Store*.

Espera-se que com a conclusão deste projeto seja possível suportar diversas soluções inovadoras para a Universidade Federal do Maranhão.

Referências

Batista, M. G. R. and Lóscio, B. F. (2013). Opensbbd: Usando linked data para publicação de dados abertos sobre o sbbd. In *SBBD (Short Papers)*, pages 10–1.

Becket, D. (2011). Re: What does sparql stand for?

Berners-Lee, T. (2006). Linked data.

Hey, T., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.

Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, 33(2):163–180.

Silva, M. L., Costa, S. S., dos Santos Oliveira, W. C., and Lima, T. M. (2017). Abrindo os dados publicos da universidade federal do maranhao. *VI Encontro Acadêmico de Computação - UFMA*.