Melhoria na Publicação de Dados Abertos: Automatização na Publicação e Indexação Semântica dos Dados

Luiz C. B. Martins[1], **Everton Agilar**[1], **Rodrigo da Fonseca Silveira**[1], **Márcio C. Victorino**[1]

[1] Centro de Informática – CPD
Universidade de Brasília (UnB) – Brasília, DF – Brazil

{luizmartins, evertonagilar, rodrigofonseca}@unb.br, mcvictorino@uol.com.br

Resumo. *A publicação de dados abertos é uma realidade no serviço público brasileiro, sendo que existe o desafio de conectar conjuntos de dados de diversas fontes agregando qualidade na publicação, neste contexto este artigo apresenta a ferramenta UnBLod (Dados Conectados da UnB) que visa automatizar a publicação de dados abertos utilizando barramento de serviços ErlangMS e realizar a indexação semântica dos dados com metadados e ontologias.*

1. Introdução

A política de abertura de dados no Brasil veio a atender aos anseios da Lei 12.257/2011 conhecida como Lei de Acesso a Informação (LAI), sancionada em 18 de novembro de 2011. A LAI foi criada com o objetivo de regulamentar o direito constitucional de todos os cidadãos terem acesso às informações governamentais de interesse público[Brasil 2011] e neste âmbito se encontra a política de publicação de dados abertos estabelecida pelo decreto Nº 8.777, de 11 de maio 2016[Brasil 2016].

A Universidade de Brasília (UnB)[1] está próxima de publicar seu Plano de Dados Abertos, onde definirá como irá implantar a política de abertura de dados, contudo, o Centro de Informática (CPD) em parceria com o Programa de Pós-Graduação em Computação Aplicada (PPCA) desenvolveu uma ferramenta que visa a otimização na publicação de dados dados abertos chamada UnBLod (Dados Conectados da UnB). Esta ferramenta gerencia a extração de dados utilizando um barramento orientado à serviço e incrementa a qualidade dos dados tornando em dados conectados por meio da indexação semântica através de ontologias e metadados. Por fim, utilizamos uma API(*Application Programming Interface*) do CKAN (*Comprehensive Knowledge Archive Network*) para que a publicação dos dados seja realizada de modo automatizadas.

Na seção 2 é apresentado os temas que envolvem a publicação de dados aberto e qual foi a estratégia e tecnologias adotadas. Na seção 3 é mostrado os resultados que foram obtidos. Por fim a seção 4 são apresentadas as conclusões, as consequências da utilização da ferramenta e possíveis sequências para este trabalho.

2. Métodos

Este processo publicação dos dados é realizado em 4 etapas: Cadastro das informações do Conjunto de Dados, Extração de Dados, Indexação Semântica e Automatização.

[1] http://www.unb.br

2.1. Cadastro de Informação do Conjunto de dados

Neste momento é selecionado o conjunto de dados para abertura que deve seguir as diretrizes definidas no plano de ação estabelecidos no Plano de Dados Abertos da instituição. Como a UnB ainda não publicou seu plano, escolhemos dados referente aos cursos e departamentos que servirão de exemplo. As informações do conjunto de dados deve ser cadastradas na ferramenta de publicação onde se preenche os metadados obrigatórios e/ou desejados[2].

2.2. Extração dos dados

A extração dos dados é realizada por meio do barramento de serviços ErlangMS desenvolvido na UnB em meados de 2014 com o objetivo de unificar o acesso aos dados da universidade através de uma camada intermediadora entre componentes de software (denominados serviços) e as aplicações que consomem estes serviços [Agilar et al. 2016].

O ErlangMS implementa o estilo arquitetural RESTful e utiliza primariamente o formato JSON para a troca de mensagens sendo que a comunicação do cliente com o barramento ocorre por meio de uma API REST padronizada desenvolvida pelo CPD/UnB. Essa API possibilita o uso de diversos tipos de operadores para facilitar a extração dos dados, tais como *filter*, *sort*, *limit* e *offset*.

Um componente chave dessa arquitetura é o conceito de catálogo de serviços que em linhas gerais, dá visibilidade aos serviços disponibilizados para extração dos dados. O catálogo de serviços contém as definições da API para acesso aos dados e os metadados para o barramento buscar os dados solicitados na extração.

Desse modo, para realizar o acesso a fonte de dados é necessário primeiramente definir a API REST do serviço no catálogos de serviços do barramento. Cada serviço é definido por um contrato onde se define qual é o escopo da consulta, a *url* do serviço, o tipo de autenticação, parâmetros de entrada que podem ser fixos (caracteres, números) ou temporais (dias, semestres, ano), entre outros atributos. Uma grande vantagem do uso barramento é a sua flexibilidade para buscar dados de diversas fontes, sejam eles um bancos de dados *SQL-Server* ou *PostgreSQL* ou de arquivos CSV(Comma-separated values) através do atributo *datasource*.

Com a definição do serviço no catálogo de serviços pronta, a extração dos dados pode ser realizada por meio de uma chamada REST ao barramento. Por exemplo, para invocar o serviço /hackathon/cursos filtrando somente os cursos do segundo semestre de 2016, seria preciso fazer uma requisição HTTP/REST no seguinte formato:

```
/hackathon/cursos?filter={ "semestre" : "20162" }
```

2.3. Indexação semântica

Tim Berners-Lee propôs um princípio que vise categorizar o nível de abertura de um dado conhecido como "5 Estrelas dos Dados Conectados" (*5 Stars Linked Data*)[Berners-Lee 2006]. Berners-Lee estabelece que a qualidade de um dado está relacionado com a sua capacidade de conexão. Assim, de acordo com o Índice 5 Estrela, um dado estruturado e em formato aberto como JSON ou CSV possuem 3 estrelas. Para

[2] `http://wiki.dados.gov.br/Padroes-de-metadados.ashx`

chegar a 4 estrelas deve obedecer padrões estabelecidos pela *World Wide Web Consortium*(W3C)[3] utilizando triplas *Resource Description Framework*(RDF) e SPARQL[4] e 5 estrelas quando é conectado a outro dado. A Figura 1 apresenta visualmente o índice 5 estrelas.

RDF é um modelo com objetivo de descrever recursos na *web* onde estes recursos possuem uma identificação única por meio de uma IRI(*Internationalized Resource Identifier*). As triplas RDF são representadas através dos itens "Sujeito" e "Objeto" e como este atores se relacionam que chamamos de "Predicados"[Bizer and Cyganiak 2014]. Na Figura 2 apresentamos um exemplo de tripla RDF.

Figura 1. 5 Estrelas dos Dados Conectados

O Sujeito da tripla sempre será um recursos, mas o Objeto pode ser um recurso ou um cadeia de caracteres que chamamos de "Literal". Quando definimos um objeto como um recurso, criamos a ligação também com as triplas que este recurso está conectado aumentando a semântica dos dados.

Antes de indexar os dados, é necessário definir padrões de metadados e ontologias que possam representar os dados semanticamente que dependerá de quais dados pretende-se publicar. No nosso exemplo, utilizamos o vocabulário *Dublin Core*(DC)[5] que é um esquema de metadados para descrever características de objetos e *Academic Institution Internal Structure Ontology* (AIISO)[6] para representação de informações universitárias.

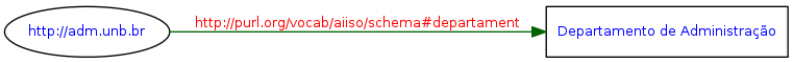

Figura 2. Tripla RDF
Fonte: Elaboração Própria

A indexação semântica é realizada no UnBLod conforme apresentado na Figura 3, Em "Definição do Sujeito", é necessário definir qual será o IRI que identificará o recurso de cada linha do conjunto de dados. O agente publicador pode informar uma URL que achar interessante e selecionar um dos campos que servirá de complemento para esta informação ou selecionar somente o campo, desde que este contenha uma URL. Na sequência será apresentado a lista do campos que o conjunto de dados fornecido pelo serviço onde se escolhe se o campo será publicado, defini o vocabulário que representará o predicado de acordo com os metadados e ontologias pre-definidos. Em "objeto" é estabelecido se é um Literal ou um outro recurso que poderá ser escolhido no campo "Complemento".

Caso exista alguma informação no conjunto de dados que não se definiu ainda

[3] https://www.w3.org/
[4] https://www.w3.org/TR/rdf-sparql-query/
[5] http://purl.org/dc/elements/1.1/
[6] http://vocab.org/aiiso/schema

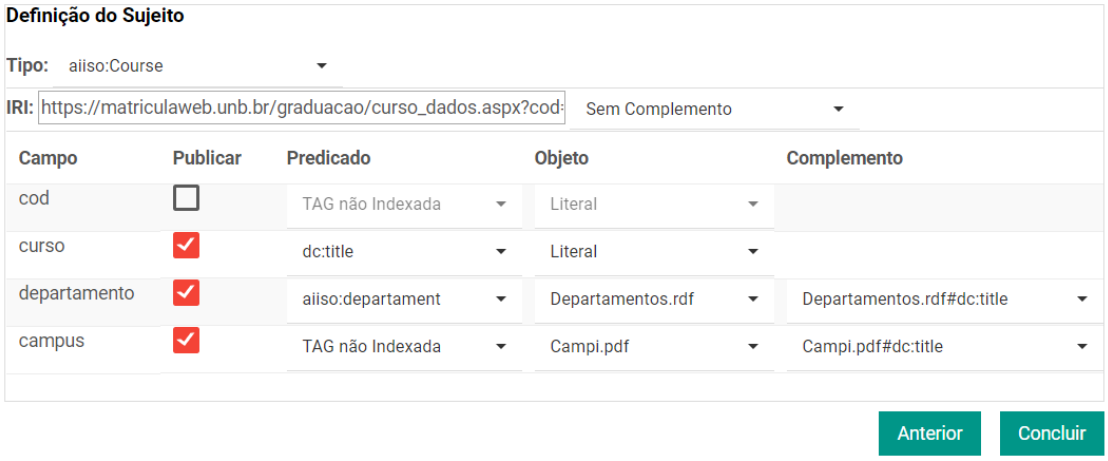

Figura 3. UnBLod - Indexação Semântica do Dados
Fonte: Elaboração Própria

um vocabulário para ela, por convenção optamos por representar através uma *tag* não indexada. Na Figura 4 é possível verificar um grafo com as triplas criadas.

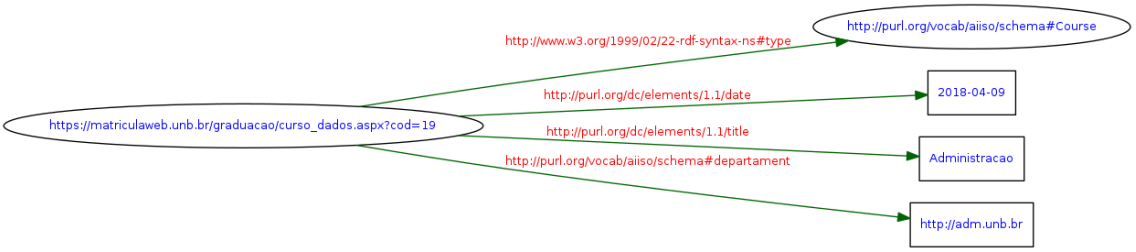

Figura 4. Grafo de Tripla RDF de Cursos
Fonte: Elaboração Própria

Destacamos aqui a tripla gerada através com predicado "http://purl.org/vocab/aiiso/schema#departament" e objeto "http://adm.unb.br": neste caso, o valor referente ao objeto aponta para outro recurso, assim se estabelece uma ligação entre os dois.

Este dados serão salvos e quando for requisitados através das solicitações configuradas pelo CKAN gerarão os arquivos CSV e RDF desde conjunto de dados.

2.4. Automatização da publicação

Para a automatização, utilizaremos as APIs do CKAN que possibilitam realizar a importação de conjuntos de dados de outras aplicações, sendo que a fonte de dados será disponibilizado pela nossa ferramenta de extração e indexação. Esse processo irá realizar nova extração e indexação semântica e fará a atualização dos dados conjuntos de dados já cadastrados. Os parâmetros definidos na seção 2.2 serão mantidos, a não ser que exista algum parâmetro temporal que será atualizado automaticamente.

3. Resultados

O UnBLod foi concebido a partir das necessidades de publicar dados aberto com mais qualidade, a principal funcionalidade desta ferramenta é estabelecer uma interface que

possibilite um agente publicador de dados informar os dados de publicação, configurar os parâmetros de acesso aos dados através do serviços disponibilizados e realizar a indexação semântica de forma a gerar o arquivo RDF conforme exemplo apresentado no Código 1.

Código 1. Exemplo de RDF.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:aiiso="http://purl.org/vocab/aiiso/schema#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
  <aiiso:Course
    rdf:about="https://matriculaweb.unb.br/graduacao/curso_dados.aspx?cod=19"
    dc:date="2018-04-09">
    <dc:title>Administracao</dc:title>
    <aiiso:departament>http://adm.unb.br</aiiso:departament>
  </aiiso:Course>
</rdf:RDF>
```

Deste modo, a ferramenta gerencia as publicações de dados abertos e possibilita o reaproveitamento de parâmetros para que possam ser utilizados em futuras atualizações. Ela faz a interface entre a fonte de dados primária e ambiente de publicação de dados aberto CKAN, garantindo a menor interferência humana e possibilitando que, através de APIs, a publicação seja automatizada.

4. Conclusão

Diante da evolução do Brasil no âmbito da abertura de dados, é importante garantir que este dados tenha a melhor qualidade possível. Hoje a maioria dos conjuntos de dados disponibilizados são publicado de forma isolada onde a integração de diferentes bases é dificultada por não haver padronização. Entre as Instituições Federais de Ensino Superior esta situação se agrava pois, teoricamente, os dados produzidos por elas possuem equivalência, portanto, podem ser ligados garantido melhor aproveitamento na abertura dos dados. Assim, apresentamos o UnBLod: uma ferramenta que auxilia na publicação dos dados conectados além de possibilitar a automatização desta atividade, dando agilidade e qualidade neste processo agregando valor e qualidade aos dados através do enriquecimento semântico.

Para trabalhos futuros, iremos implementar uma interface de consulta nos nossos arquivos utilizando SPARQL onde as resultados serão enriquecidos semanticamente e também oferecer esta interface que outras instituições que desejam indexar seus dados semanticamente possam realizá-los de maneira mais eficiente.

Referências

Agilar, E., Almeida, R., and Canedo, E. (2016). A systematic mapping study on legacy system modernization. In *SEKE*.

Berners-Lee, T. (2006). Desing Issues - Linked Data.

Bizer, C. and Cyganiak, R. (2014). Rdf 1.1 TriG-RDF dataset language-w3c.

Brasil (2011). LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011.

Brasil (2016). Decreto nº 8.638 de 15, de janeiro de 2016.