WEB SEMÂNTICA E EXTRAÇÃO DE DADOS NA COMPOSIÇÃO DE MODELO ESTRUTURAL PARA DADOS DE RESULTADOS DE PRODUÇÃO CIENTÍFICA

José Eduardo Santarem Segundo
Universidade de São Paulo, São Paulo, SP, Brasil
Universidade Estadual Paulista, São Paulo, SP, Brasil
santarem@usp.br

Dalton Lopes Martins
Universidade de Brasília, Brasília, DF, Brasil
dmartins@gmail.com

1 INTRODUÇÃO

Inicia-se a apresentação desta pesquisa contextualizando-se o cenário que é objeto deste estudo. Trata-se de uma pesquisa que envolve: comunicação científica, mais propriamente um recorte de dados de resultados de pesquisa científica, caracterizados por publicações em anais de eventos que utilizam a ferramenta *Open Conference System* (OCS/PKP), assim como todo o contexto que os envolvem, como autores, assunto, vínculos institucionais e citações; metodologias e processos que envolvem conceitos de dados abertos, web semântica e ligação de dados (*Linked Data*); e principalmente técnicas de extração e mineração de dados, baseados em *Web Scraping*.

A pesquisa aqui apresentada foi motivada pela análise das metodologias e técnicas utilizadas para a extração e acesso a dados, pelos pesquisadores das áreas de Bibliometria e Cientometria, em pesquisas apresentadas nos últimos (3 ocorrências) Encontros Brasileiros de Bibliometria e Cientometria (EBBCs) e no Grupo de Trabalho 7 – GT7 (Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação) do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB).

Ao analisar os métodos de extração e acesso a dados, é notável identificar que o processo de acesso e organização dos dados é muitas vezes realizado manualmente, com falta de recursos computacionais adequados e/ou usando técnicas que colocam em risco a credibilidade dos resultados apresentados posteriormente.

Apesar de esta pesquisa tratar com ênfase de processos que envolvem Web Semântica, ela não se inicia na organização semântica e ligação de dados por produtores de conteúdo, ou por editores de revistas ou ainda coordenadores de eventos científicos, o que seria formidável para alavancar estudos cientométricos, ela parte justamente de uma necessidade por esses processos não acontecerem. Nesse caso apresenta-se aqui uma proposta inicialmente baseada em extração de dados, que posteriormente é submetida a processos que envolvem mapeamento e enriquecimento de dados ligados semanticamente.

Entende-se, portanto, que a utilização de técnicas computacionais de extração de dados e principalmente de ligação de dados, através das melhores práticas de *Linked Data* e Web Semântica, pode contribuir sobremaneira para o avanço de estudos bibliométricos e cientométricos.

Esta pesquisa tem como objetivo principal constituir um modelo de aplicação que favoreça a recuperação e acesso a dados de resultados de produção científica, por meio de um *framework*, baseado em técnicas e algoritmos de extração e mineração de dados e conceitos e técnicas de Web Semântica e das melhores práticas de *Linked Data*. Acredita-se que esse modelo de aplicação pode estimular novos estudos bibliométricos e cientométricos, com base em dados originários de anais de eventos que utilizam a ferramenta OCS/PKP.

Para atingir os objetivos deste trabalho, utilizou-se uma metodologia qualitativa exploratória, por meio de literatura técnica e científica que embasam os conceitos de extração e mineração de dados e textos, Web Semântica e melhores práticas de *Linked Data* e o uso de ferramental para implementação de modelo de aplicação.

Destaca-se e justifica-se o uso de fonte de dados baseada em anais de eventos disponibilizados por meio da ferramenta OCS/PKP.

Apesar da opção por esta escolha parecer um tanto redutora, o OCS/PKP é uma das ferramentas mais utilizadas em todo o mundo para gerenciamento de conferências, tornando o modelo proposto aqui uma ótima referência e, portanto, aceito para análises de dados de um grande número de conferências. Justifica-se ainda que todo o modelo proposto encontra-se em fase evolutiva, e que as técnicas de extração encontram-se em fase de desenvolvimento, e que evoluirão para o uso de algoritmos de *Machine Learning* (aprendizagem de máquina), permitindo a expansão a outros tipos de ferramenta (além do OCS/PKP) que disponibilize os dados em formato aberto.

Pretende-se com a apresentação desse modelo de aplicação, por meio de relações constituídas semanticamente, disponibilizar para a comunidade científica, pesquisadores, autores e outros interessados no tema, um ferramental que permita com que com poucos cliques seja possível responder questões como: número de artigos por autor, relações de coautoria entre autores e institucionais, relações dentro e fora de uma instituição, autores por assunto, e numa versão mais ampliada autores mais citados individualmente ou por seção e outras relações de citação, entre tantas outras perguntas que são pertinentes quando se trata de dados de resultados de produção científica.

2 EXTRAÇÃO DE DADOS

Os processos de extração e mineração de dados vêm se desenvolvendo em tecnologia estatística desde o século XIX. A mineração de dados tem ganhado cada vez mais força, principalmente com o fortalecimento do chamado quarto paradigma da ciência, *e-Science* ou ainda *Data-Driven Science*, que entende os dados como um grande aliado e impulsionador para o avanço da ciência moderna. Como previsto por Jim Gray em 2007, o quarto paradigma da ciência se faz cada dia mais presente nas ações do mundo atual.

As técnicas de extração de dados nesta pesquisa são algoritmos de *Web Scraping*, que são coletas automatizadas de dados da Internet. Apesar de não ser uma técnica ou termo novo, a prática já fora chamada de

data mining, *web harvesting*, *screen scraping*, entre outras. Atualmente, há um consenso sobre o uso do termo *Web Scraping* para se referir a técnicas de coletar dados de páginas web por meio de uma *Application Program Interface* (API) ou acesso humano (MITCHELL, 2016). O processo de *Web Scraping* consiste em desenvolver um programa que seja capaz de consultar um servidor Web, solicitar dados (como se fosse uma página WEB) e extrair informações ao analisá-los.

Realizar um *scraping* em dados de páginas Web possibilita acessar dados via um *script* de programação, organizá-los, armazená-los em bancos de dados e posteriormente fazer qualquer tipo de análise.

3 WEB SEMÂNTICA E *LINKED DATA*

A Web Semântica, proposta por Tim Berners-Lee, James Hendler e Ora Lassila, em artigo publicado na *Scientific American* em 2001 propõe, essencialmente, que se estruture os dados da Web de forma que eles possam ter significado e principalmente que se tornem passíveis de interpretação por máquinas, através de agentes computacionais. Os autores destacam as linguagens *XML* (atualmente JSON tem assumido este papel) e *RDF* como essenciais para a consolidação da Web Semântica, assim como definem as ontologias como responsáveis por organizar o conhecimento neste novo paradigma da Web.

Por meio do uso das tecnologias citadas, caracterizam-se as ontologias como um dos principais elementos da Web Semântica na construção de informações relacionadas que apresentem significado. Santarem (2015, p. 226) afirma que:

> Utilizar ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação.

A OWL é uma linguagem de marcação semântica para a definição, a instanciação, a publicação e a partilha de ontologias na *World Wide Web*. A

linguagem *OWL* é reconhecida, atualmente, como o último padrão em linguagens para ontologia e é recomendada como a principal linguagem para construção de ontologias pelo consórcio W3C.

Um caminho que floresce nesta revolução se encontra no *Linked Data*, um cenário vislumbrado por Tim Berners-Lee, que a partir dos conceitos e das tecnologias da Web Semântica busca ser uma forma de publicar dados na Web. O *Linked Data* contempla essencialmente diretrizes para a disponibilização de dados na Web, interligando variados conjuntos de dados, seguindo os princípios da Web de Dados.

A chamada Web de Dados faz com que dentro da Web, as informações estejam contextualizadas e relacionadas a outros recursos, possibilitando que agentes computacionais sejam capazes de compreender o domínio de um dado, aprimorando a recuperação da informação. O *Linked Data*, criado em 2006 por Berners-Lee, é uma proposta que utiliza os conceitos da Web Semântica e se destacou pela inserção de significado nos dados na Web.

De acordo com Berners-Lee (2006), o conceito está relacionado ao uso de ferramentas para ligação de dados da seguinte forma:

> *Linked Data* é um corpo crescente de conjuntos de dados na rede mundial que estão interligados por meio do recurso *Description Framework* (RDF) usando identificadores de recursos uniformes baseados na web (URIs) para identificar tanto as coisas descritas e os termos usados para descrevê-los.

Para Hooland e Verborgh (2014), o *Linked Data* não pode ser encarado como uma ferramenta, mas como melhores práticas para estruturação de dados.

4 MODELO DE APLICAÇÃO

O modelo de aplicação proposto baseia-se na extração de dados do recurso de publicação de anais da ferramenta OCS/PKP, que disponibiliza seus dados em uma página formatada em HTML. Nesses resultados são informados o título do trabalho, o nome dos autores e um *link* tanto

para o acesso ao resumo (numa segunda página) quanto para o arquivo depositado (normalmente em PDF).

Como é possível observar por meio da Figura 1, o acesso a essa segunda página permite identificar os mesmos dados da página geral, e ainda as informações do resumo, as palavras-chave, a instituição dos autores (quando informado nos metadados do sistema), a agência de fomento, o nome da conferência e o agrupamento (no caso do exemplo o GT1 do ENANCIB).

FIGURA 1 - CÓDIGO FONTE DE PÁGINA DE PUBLICAÇÃO EM ANAIS OCS/PKP, ENANCIB - 2017

```
<meta name="DC.Source" content="XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XVIII ENANCIB)"/>
<meta name="DC.Source.URI" content="http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/index"/>
                    <meta name="DC.Subject" xml:lang="pt" content="Documento"/>
                                <meta name="DC.Subject" xml:lang="pt" content="Dispositivo"/>
                                <meta name="DC.Subject" xml:lang="pt" content="Práticas Documentárias"/>
                                <meta name="DC.Subject" xml:lang="pt" content="Efeitos de informação"/>
            <meta name="DC.Title" content="DOCUMENTO E DISPOSITIVO: ENTRE BERND FROHMANN E MICHEL FOUCAULT"/>
        <meta name="DC.Type" content="Text.Proceedings"/>
    <meta name="DC.Type.paperType" content="GT-1 – Estudos Históricos e Epistemológicos da Ciência da Informação - Comunicação Oral"/>          <meta name="gs_meta_revision"
content="1.1" />
    <meta name="citation_conference_title" content="XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XVIII ENANCIB)"/>
    <meta name="citation_author" content="Thays Lacerda Ferrando"/>
    <meta name="citation_author_institution" content="Universidade Federal Fluminense"/>
    <meta name="citation_author" content="Lidia Silva de Freitas"/>
    <meta name="citation_author_institution" content="Universidade Federal Fluminense"/>
    <meta name="citation_title" content="DOCUMENTO E DISPOSITIVO: ENTRE BERND FROHMANN E MICHEL FOUCAULT"/>
    <meta name="citation_date" content="2017/08/19"/>
    <meta name="citation_abstract_html_url" content="http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/542"/>
    <meta name="citation_pdf_url" content="http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/download/542/1079"/>

    <link rel="stylesheet" href="http://enancib.marilia.unesp.br/lib/pkp/styles/common.css" type="text/css" />
    <link rel="stylesheet" href="http://enancib.marilia.unesp.br/styles/common.css" type="text/css" />
    <link rel="stylesheet" href="http://enancib.marilia.unesp.br/styles/paperView.css" type="text/css" />
```

Fonte: Dados da pesquisa, 2018.

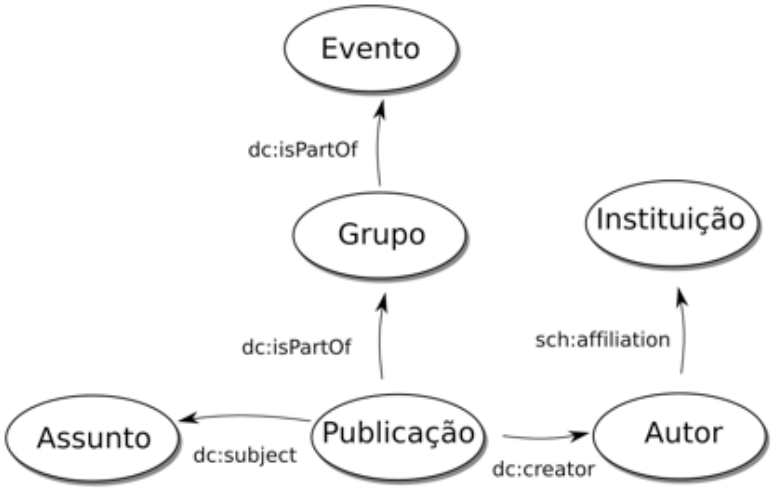
Desenvolveu-se um algoritmo na linguagem *Python* que executa a leitura completa do documento e gera um arquivo no formato *Turtle* (linguagem para geração de código formalizado de RDF) com todas essas informações, tornando cada uma dessas publicações um recurso com suas devidas propriedades RDF. O início de todo o processo de extração e geração completa dos dados é realizado pela inserção do *link* que compõe os anais no OCS/PKP. O algoritmo em *Python* realiza ainda uma indexação de termos (palavras-chave) e também de autores, identificando quando esses elementos se repetem em trabalhos distintos.

A seguir é feito o processo de enriquecimento dos dados. Neste caso específico, ele é realizado a título de exemplo com uma técnica de associar os termos encontrados a conceitos no *DbPedia* e também com a execução de um código de extração que realizada uma busca do Número ORCID do autor na base ORCID, criando o vinculo de enriquecimento.

O modelo aqui apresentado, que ainda evoluirá, conta com uma ontologia que tem como ponto de partida a própria publicação. A partir da publicação geram-se as ligações para os autores e suas respectivas instituições, as palavras-chave (assunto), os grupos que as publicações pertencem (quando isso ocorre dentro do evento) e ainda uma relação de informação com a conferência consultada. O modelo pode ser utilizado para integração de dados de várias conferências.

Por meio da Figura 2 é possível identificar a estrutura formal que dá origem a ontologia gerada para esse projeto.

FIGURA 2 - ESTRUTURA FORMAL DE DADOS



Fonte: Dados da pesquisa, 2018.

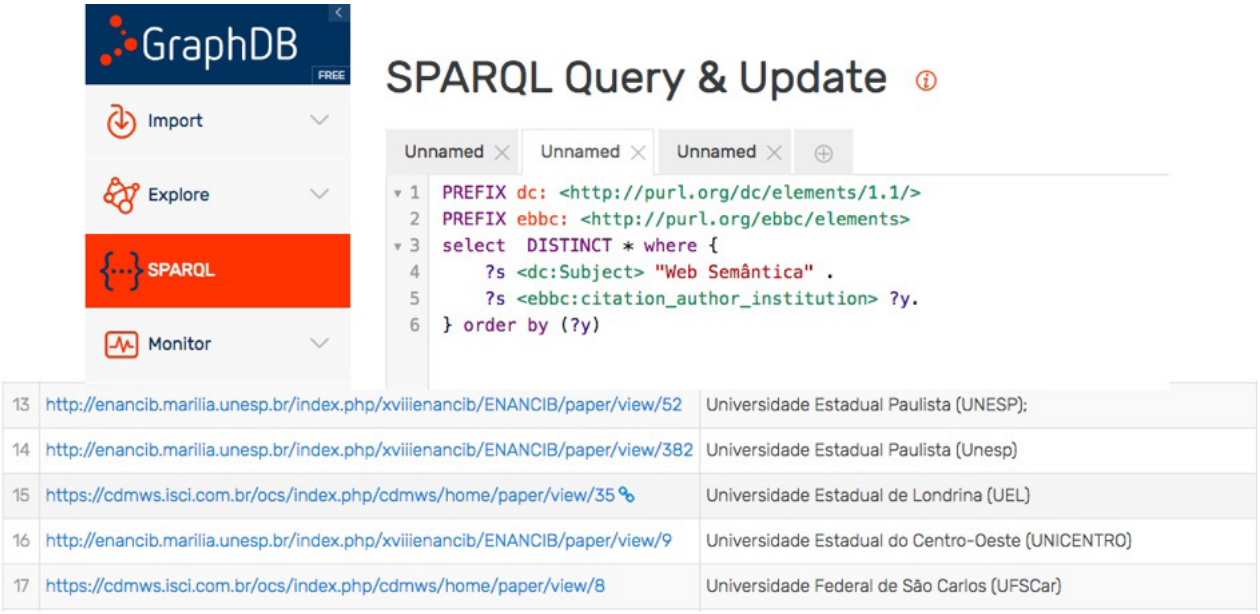
A partir do modelo formal gera-se uma estrutura computacional utilizando-se linguagem RDF e aplica-se inserção do conjunto de dados, já organizados e enriquecidos, em um banco de dados de triplas.

O banco de dados de triplas, disponibiliza uma interface conhecida como *SparqlEndPoint*, que apesar de ainda ser uma estrutura um tanto complicada para o uso de leigos, é uma ferramenta que permite realizar um grande conjunto de combinações para recuperar informações como as propostas no começo deste texto.

Uma prova de conceito foi realizada utilizando-se os anais do Enancib 2017 e do Colóquio de Dados, Metadados e Web Semântica. Nesta prova de conceito procedeu-se a extração e enriquecimento dos dados, assim

como a alimentação de uma base de dados de triplas utilizando-se o software *GraphDB*. Com a agregação desses dados e disponibilização para consultas *Sparql* foi possível realizar buscas de vários tipos, por meio da Figura 3 é possível verificar o código *Sparql* de consulta e parte do resultado dessa consulta que busca entre informações de publicações dos dois congressos, recuperando as instituições que tiveram trabalhos publicados com a palavra-chave "Web Semântica".

FIGURA 3 - CONSULTA E RESULTADO *SPARQL*, USANDO O TERMO WEB SEMÂNTICA.



Fonte: Autores, 2018.

O resultado apresentado, possível de ser visto no recorte por meio da Figura 3, mostra os links dos recursos (trabalhos), com as instituições a qual os autores são filiados. Nota-se que há recursos referentes aos dois congressos que foram utilizados como base.

5 CONSIDERAÇÕES FINAIS

Não há dúvidas que a evolução de estudos cientométricos e bibliométricos passa necessariamente pela evolução das tecnologias no processo de extrair e mapear dados oriundos dos mais diversos tipos de fontes de publicação de dados científicos.

O modelo apresentado aqui propõe um cenário animador para os avanços na disponibilização de dados e acesso a informações de publicações em eventos.

Apesar do modelo ainda ter muito a evoluir com uso de *Machine Learning* (Aprendizagem de Máquinas), alimentação da base com um novo processo de extração das citações, construção de axiomas que permitam inferências, geração de um protótipo disponível para a comunidade, e ainda uma infinidade de outras técnicas, os recursos e as técnicas de extração e melhores práticas de *Linked Data* utilizados, já nos permitiram chegar a resultados que acelerarão sobremaneira os estudos dos dados citados.

Destaca-se que o processo de extração, o enriquecimento e a população de uma base de dados de triplas, que dura aproximadamente cinco minutos (teste realizado com as publicações do Enancib de 2017 e do Colóquio de Dados, Metadados e Web Semântica, também de 2017), transforma todo o conteúdo disponibilizado em anais de eventos (OCS/PKP) em uma base de dados semântica, estruturada e passível de variadas consultas.

Não há como não observar que todo o processo de geração de uma base semântica poderia ser realizado pela própria ferramenta OCS/PKP, ou pela coordenação dos eventos, entretanto, na falta deste consideramos totalmente satisfatório e útil realizar a primeira parte do processo com extração de dados.

Considera-se importante notar que uma base de dados organizada semanticamente dá margem a uma infinidade de tipos de consultas, que envolvam todos os tipos informações que compões a ontologia, como: autores, instituições, palavras-chave, eventos e suas segmentações de publicações, agências de fomento, entre outras informações que possam vir advindas dos processos de enriquecimento.

REFERÊNCIAS

BERNERS-LEE T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, New York, v. 5, May 2001.

BERNERS-LEE, T. **Linked Data**: Design Issues 2006. Disponível em <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 10 jun. 2016.

HOOLAND, S. van; VERBORGH, R. **Linked Data for libraries, archives and museums**: how to clean, link and publish your metadata. London: Facet, 2014.

MITCHELL, R. **Web Scraping com Python**. São Paulo: Novatec, 2016.

SANTAREM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 8, p. 219-239, 2015. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207.> Acesso em: dez. 2017.