Anotação semântica automática do currículo Lattes utilizando Linked Open Data

Walison Dias da Silva

Mestre em Sistemas de Informação e Gestão do Conhecimento Universidade Fumec.

Fernando Silva Parreiras

Doutor em Ciência da Computação. Professor do programa de mestrado e doutorado em Sistemas de Informação e Gestão do Conhecimento da Universidade Fumec.

Luiz Cláudio Gomes Maia

Doutor em Ciência da Informação. Professor do programa de mestrado e doutorado em Sistemas de Informação e Gestão do Conhecimento da Universidade Fumec.

Wladmir Cardoso Brandão

Doutor em Ciência da Computação pela UFMG. Professor adjunto da PUC-MINAS.

http://dx.doi.org/10.1590/1981-5344/3185

A Web semântica possui a finalidade de otimizar a recuperação dos documentos, sendo que estes recebem significados, permitindo que tanto as pessoas quanto as máquinas possam compreender o significado de uma informação. A anotação semântica de entidades é o caminho para promover mais semântica em documentos. O objetivo do artigo é propor uma abordagem que possibilite anotar automaticamente entidades nos Currículo Lattes de pesquisadores por meio de bases de dados abertas (Linked Open Data), as quais armazenam o significado de termos e expressões. O problema da pesquisa está baseado em saber quais são os conceitos associados à Web Semântica que podem contribuir para a Anotação Semântica Automática do Currículo, Lattes utilizando o Linked Open Data (LOD). Na pesquisa foram apresentados conceitos, ferramentas e tecnologias relativas ao tema. A aplicação destes conceitos possibilitou a criação do Sistema Lattes Web Semântico. Um experimento empírico foi realizado com o objetivo de

identificar a ferramenta de extração de entidade mais efetiva. O sistema possibilita a importação do currículo XML da Plataforma Lattes, efetua a anotação automática dos dados disponibilizados utilizando as bases de dados abertas e possibilita efetuar consultas semânticas.

Palavras-chaves: *Anotação Semântica. Dados Abertos Interligados. Plataforma Lattes.*

Semantic Annotation Automatic of Curriculum Lattes Using Linked Open Data

The Semantic Web has the purpose of optimizing document recovery, where these documents received synonyms, allowing people and machines to understand the meaning of one information. The semantic annotation entity is the path to promote the semantic in documents. This paper has an objective to build an outline with the Semantic Web concepts that allow to automatically annotate entities in the Lattes Curriculum based on Linked Open Data (LOD), which store terms and expressions' meaning. The problem addressed in this research is based on what of the Semantic Web concepts can contribute to the Automatic Semantic Annotation Entities of the Lattes Curriculum using Linked Open Data. During the literature review the concepts, tools and technologies related to the theme were presented. The application of these concepts allowed the creation of the Semantic Web Lattes System. An empirical study was conducted with the objective of identifying an Extraction Tool Entity further Effective. The system allows importing the XML curricula in the Lattes Platform, annotates automatically the available data using the open databases and allows to run semantic queries.

Keywords: *Semantic Annotation Entity. Linked Open Data. Lattes Platform.*

Recebido em 18.06.2018 Aceito em 11.09.2018

1 Introdução

A Web Semântica (WS), proposta em 2001 por Berners-Lee et al. (2001), como uma extensão da Web atual, onde as informações possuem

significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação. Ela surge com o propósito de solucionar o problema de recuperação de dados, interoperabilidade e compartilhamento de conhecimento, em que as informações são atribuídas (anotadas) a seus significados, permitindo que tanto as pessoas quanto as máquinas, possam compreender o significado de uma informação. Com a web semântica a Internet passa a funcionar de outra forma, pois em uma rede de informações, cada item passa a conter o seu significado, o que permite melhores interações com o usuário. Com o objetivo de criar um documento na web tradicional (WT) que seja passível de interpretação humana, seja analisado e processado por máquinas e softwares de modo a realizarem pesquisas precisas surge proposta da Anotação Semântica (AS) (FONTES et al., 2010). Diferente da WT, onde os documentos se relacionam utilizando *links* sem significado definido, essa proposta transmite as palavras significados que viabilizam sistemas de buscas precisos. Isso implica na melhoria de resultados de pesquisas na Web, pois não será necessário procurar uma determinada informação em uma série de páginas de resultados genéricos, já que serão exibidas páginas que definem a palavra procurada.

A Plataforma Lattes (PL) representa a integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um Sistema de Informações. Sua dimensão atual se estende às ações de planejamento, gestão e operacionalização do fomento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Estátisticas do Currículo Lattes apomtam 531 instituições, registrando 37.640 grupos e 199.566 pesquisadores, sendo 129.929 doutores[1]. Além disso, a plataforma Lattes se tornou estratégica para as atividades de planejamento, gestão, para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação (CNPQ, 2014).

A Anotação Semântica é um esquema para geração e uso de metadados, habilitando novos métodos de acesso a informação. Pesquisas na área da anotação semântica são relevantes para solucionar problemas de busca, de localização e de recuperação da informação.

Este artigo tem como objetivo propor um arcabouço para anotação automática de entidades no Currículo Lattes de pesquisadores por meio das ligações de bases abertas (*Linked Open Data*). O trabalho apresentará conceitos, ferramentas, tecnologias que venham enriquecer os dados do Lattes semanticamente e anotar automaticamente por intermédio dos dados interligados. O presente artigo contribui com a pesquisa na envolvendo Web Semântica e a PL, dá sequência e explora limitações de estudos parecidos (BONIFACIO, 2002; CASTANO, 2008; GALEGO, 2013). O propósito é permitir que os dados sejam encontrados e indexados na Web, aberto e disponível em formato compreensível por máquina, além de

[1] Disponível em: <http://lattes.cnpq.br/web/plataforma-lattes/dados-e-estatisticas>. Acesso em: 30 jan. 2016.

estar disponível para ser reaplicado em outros sistemas e domínios (GOV, 2014). Não tem como objetivo aprofundar exaustivamente na análise da eficiência da ferramenta de anotação semântica.

Este artigo foi estruturado da seguinte maneira em introdução, referêncial teórico trabalhando o os conceitos relacionados a Web Semântica, características e ferramentas para Anotação Semântica, onde são abordados trabalhos já desenvolvidos do Lattes relacionado com o tema proposto, metodologia com a apresentação do arcabouço de anotação semântica automática utilizando os dados abertos conectados (*Linked Open Data)* e as considerações finais.

O trabalho é destinado a pesquisadores que exploram o tema da Web Semântica e procuram adquirir conceitos e práticas relacionadas com a anotação semântica, extrator de entidade, *linked open data*, Resource Description Framework (RDF), *Resource Description Framework in Attributes* (RDFa) e SPARQL.

2 Referencial teórico

2.1 Fundamentos e conceitos da web semântica

Temos as principais tecnologias e camadas relacionadas ao projeto da Web Semântica:

***LINKED DATA* (dados conectados):** é um conjunto de boas práticas para publicar e conectar dados estruturados na *Web* (BERNERS-LEE et al*.,* 2001). O *Linked Data* (LD) diz respeito aos dados disponíveis na *Internet* que são compreendidos por máquinas assim como pelas pessoas, com significado definido, ligado a outros conjuntos de dados externos e que por sua vez, está conectado a outro conjunto de dados (CHRISTIAN BIZER, 2009). *Linked Data* em 2014 apresentou um conjunto de 570 dados ligados à *Web* e 2.909 relações de ligações entre esses conjuntos (PLANETDATA, 2014). Como exemplo da aplicação dos princípios dos dados ligados (LD), temos o projeto *Linking Open Data* (LOD) onde podemos destacar a DBpedia.

RDF (*Resource Description Framework*): é um modelo de dados que cria declarações no formato de triplas (sujeito, predicado, objeto), possibilitando a descrição dos recursos por meio de suas propriedades e valores. O *Resource Description Framework* (RDF) foi proposto como uma solução para a limitação da XML que não possibilita descrever a semântica de uma informação. O RDF pode ser representado ou serializado nos seguintes formatos: RDF JSON-LD, RDF N-Quads, RDF TriG, RDF TURTLE (ttl), RDF N3 (N-Triples) e RDFa (PRIMER, 2014; CEWEB, 2016).

RDF Schema: é uma extensão da RDF que permite a definição de esquemas para os vocabulários (termos) utilizados nas declarações. É uma linguagem que permite a construção de ontologias com expressividade e inferência, pois fornece um conjunto básico de elementos para a modelagem, e poucos desses elementos podem ser utilizados para inferência. O RDFs é uma extensão semântica do RDF, que fornece

maneiras para descrever grupos de recursos e as relações entre esses recursos. Esses recursos são utilizados para especificar as características de outros recursos, como domínios e faixas de propriedades (WORLD WIDE WEB CONSORTIUM - W3C, 2014).

RDFa *(Resource Description Framework in Attributes)*: tem por objetivo embutir código RDF em estruturas HTML e XML realizado através da inclusão de significado via atributos dos elementos. A vantagem da utilização do RDFa é que máquinas de buscas podem melhorar seus resultados aumentando a precisão sobre o real significado de um documento. Ou seja, as máquinas de buscas podem agregar os dados de um documento com dados de outro documento, enriquecendo os resultados de buscas (PRIMER, 2014; CEWEB, 2016).

Ainda sobre RDFa e de acordo com CEWEB (2016) existem quatro atributos *(Prefix, Resource, Property e Typeof)* com o objetivo de descrever código RDF:

1. ***Prefix*** tem o objetivo descrever os vocabulários que estão sendo reusados no documento HTML. Por exemplo, os vocabulários FOAF, Schema.org e Dublin Core;
2. ***Resource***, cujo objetivo é descrever os recursos;
3. ***Property***, cujo objetivo é descrever uma propriedade. Uma propriedade tem o objetivo de relacionar dois elementos, ou seja, relacionar um sujeito à um objeto;
4. ***Typeof***, que equivale ao atributo para representar o elemento *rdf:type* (tem o mesmo objetivo do *rdf:type*).

RDF TURTLE (ttl): Criado para possibilitar descrever prefixos e IRIs (*Internationalized Resource Identifier*) relativos na estrutura do documento. Com a figura 1 observarmos que as seis primeiras linhas do código mostram os IRIs que podem ser definidos como prefixos e IRI base do documento, característica esta não permitida no N-Triples. Podemos observar que as linhas 08, 14 e 18 apresentam sujeitos com seus predicados e objetos logo abaixo deles. Este tipo de organização e endentação torna intuitiva a leitura do documento, facilitando a identificação das triplas RDF. O elemento que é responsável por relacionar o sujeito ao predicado deste exemplo é o *token* "**a"**. Este *token* possui a mesma semântica da propriedade **rdf:type** e é usada para dizer que bob#me é do tipo foaf:Person (PRIMER, 2014; CEWEB, 2016).

Figura 1 – Representação de Triplas RDF Turtle(TTL)

```
01    BASE   <http://example.org/>
02    PREFIX foaf: <http://xmlns.com/foaf/0.1/>
03    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
04    PREFIX schema: <http://schema.org/>
05    PREFIX dcterms: <http://purl.org/dc/terms/>
06    PREFIX wd: <http://www.wikidata.org/entity/>
07
08    <bob#me>
09        a foaf:Person ;
10        foaf:knows <alice#me> ;
11        schema:birthDate "1990-07-04"^^xsd:date ;
12        foaf:topic_interest wd:Q12418 .
13
14    wd:Q12418
15        dcterms:title "Mona Lisa" ;
16        dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .
17
18    <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
19        dcterms:subject wd:Q12418 .
```

Fonte: PRIMER (2014).

SPARQL *(Protocol and RDF Query Language)*: uma linguagem de consulta e protocolo de acesso a dados em RDF. Utilizada na recuperação de informações em aplicações da Web Semântica. Os resultados da anotação são salvos no formato *Resource Description Framework* (RDF), que fornece um modo padrão para compartilhamento de dados, intercâmbio, permite consultar e manipular os dados usando a linguagem de consulta SPARQL (TAO et al., 2013). Assim como, a linguagem SQL está para os bancos de dados relacionais, também está o SPARQL para os bancos de dados de triplas RDF. Entretanto, consultas que necessitam relacionar diferentes dados tendem a ser bem complexas em SQL do que em SPARQL (CASTANO, 2008).

A estrutura de uma consulta SPARQL é definido pelos comandos ***PREFIX*** onde declaramos os *namespaces* utilizados na consulta, ***SELECT*** declara-se o conjunto de resultados almejados, em ***FROM*** declara-se o conjunto de dados a serem consultados (os grafos RDF que serão consultados), na cláusula ***WHERE*** monta-se a condição de triplas que o resultado da pesquisa deverá satisfazer e ***ORDER BY, DISTINCT, LIMIT*,** entre outros são os modificadores da consulta.

As tecnologias descritas acima, como o RDF e o SPARQL são padrões criados pelo *Internet Engineering Task Force* (IETF), e estão naturalmente articuladas para compor os padrões da Web Semântica.

2.2. Anotação semântica

Anotação semântica é uma abordagem para alcançar os conceitos da *Web* semântica, cuja organização de informações fornece um meio, por onde a conexão lógica dos termos estabelece interoperabilidade entre sistemas. Ela é um esquema para geração e uso de metadados, habilitando novos métodos de acesso a informação. Uma anotação semântica é uma associação entre as expressões ou termos relevantes de um documento e os conceitos descritos em uma ontologia (BELLOZE et al*.*, 2012).

São propostas para realizar anotação semântica as linguagens RDFa, Micro formatos e Microdata. Todas se caracterizam por utilizar um conjunto de atributos oriundos de um vocabulário, marcando trechos de um documento HTML ou xHTML, através de triplas semelhantes às utilizadas em RDF (FONTES et al., 2010). Porém o RDFa possui os benefícios do RDF, possuindo recomendação pela W3C para interoperabilidade e legibilidade dos dados, é considerado flexível e semanticamente melhor que os microformatos.

Segundo Eller (2008), uma anotação semântica de um documento descreve o seu conteúdo pela associação de palavras relevantes do texto e conceitos presentes na ontologia. O resultado de uma anotação A é uma tupla (as, ap, ao, ac), onde: **as** é o dado anotado; **ao** é a anotação em si; **ap** é o predicado que define o tipo de relacionamento entre o **as** e **ao**; **ac** é o contexto em que a anotação foi feita (OREN et al*.*, 2006).

A anotação semântica (A.S) ou uma plataforma de anotação semântica (PAS) pode ser caracterizada e classificada pela:

1. **Forma de anotar:** Tipo manual, automática e híbrida;
2. **Forma de gravar as anotações:** Forma intrusiva e não intrusiva;
3. **Forma de descobrir entidades:** Baseado em padrões e em aprendizagem de máquina;
4. **Arquitetura:** Extensiva e não extensiva.

A maioria das plataformas de anotação semântica (PAS) lidam com um sistema de extração de informação (IE) externa, das quais a maioria foi desenvolvida a partir da comunidade de processamento de linguagem natural (PLN). Alguns sistemas de IE dispõem de serviços adicionais, tais como reconhecimento de entidades nomeadas (NER). De acordo com Derczynski et al. (2014), o reconhecimento de entidades mencionadas (NER) é uma tarefa crítica na extração de informação (IE), uma vez que identifica quais trechos do texto são menções de entidades no mundo real. De acordo com Reeve e Han (2005), as Plataformas de Anotação Semântica (PAS) podem ser distinguidas pelo seu método de anotação, sendo esse o componente que tem a maior relevância sobre à eficácia de uma anotação semântica.

2.3. Plataforma lattes e a web semântica

O Lattes é a base de dados de currículos, instituições e grupos de pesquisa das áreas de Ciência e Tecnologia. É um padrão nacional da vida pregressa e atual dos pesquisadores e estudantes. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, tornou-se um elemento essencial à análise de mérito e competência das solicitações de financiamentos na área de ciência e tecnologia.

É um sistema estratégico para as atividades de planejamento e gestão. É utilizado na formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação.

A plataforma Lattes (PL) é fonte de estudo para trabalhos como o de Bonifacio (2002) e Galego (2013) que desenvolveram trabalhos relacionadas com a Web Semântica (criação de ontologias e consultas semânticas), mas não foi encontrado pesquisas sobre a A.S. utilizando LOD. Nesses trabalhos, encontramos a sugestão de Galego (2013) que é, explorar as funcionalidades de *Linked Data* para que seja possível integração com outras bases de conhecimentos., que vem ao encontro com o núcleo desta pesquisa (anotação semântica com *Linked Open Data*) (CASTANO, 2008).

No desenvolvimento dessa pesquisa foram identificados e analisados trabalhos que associam à Web Semântica com a Plataforma Lattes (PL), dentre os quais temos:

1. O trabalho desenvolvido por Bonifacio (2002 introduz conceitos básicos sobre paradigmas de linguagens e ferramentas que estão dando os primeiros passos em direção a Web Semântica na PL. Um processo de tradução semi-automática dos dados do documento XML gerado pela exportação do Sistema Lattes para o modelo ontológico em DAML+OIL foi apresentado, sendo que a contribuição desenvolvida nesse trabalho foi permitir uma melhor compreensão dos conceitos, linguagens e ferramentas que foram apresentadas, com a aplicação deles no caso do Currículo Lattes. Como resultado, foi elaborado uma proposta de uma ontologia para a plataforma na linguagem DAML+OIL, denominada de OntoLattes.
2. Castano (2008) utilizou como fonte de informações os currículos da PL para popular automaticamente uma ontologia e utilizá-la como uma base de dados a ser consultada para geração de relatórios. Todo processo de extração de informações (wrappers) foi executado a partir de documentos HTML, com processamento posterior para inserção correta na ontologia, de acordo com sua semântica. Nesse processo, foram encontrados dificuldades ou problemas como: identificar corretamente os textos dos arquivos originais para que fosse possível mapear a ontologia com a semântica correta dos termos, identificar e retirar as duplicidades de instâncias que se referem a um mesmo objeto. No trabalho foram utilizadas duas abordagens na busca por similaridades e demonstrado suas características principais. Também foi exemplificado, de forma superficial, uma comparação da criação de consultas em SPARQL, XQuery e SQL.
3. O trabalho de Galego, (2013), foi apresentado em uma revisão de trabalhos que propuseram a geração de relatórios sumarizados de um grupo de pessoas, alguns com desenvolvimento de ontologias, no domínio do Lattes. Ele descreve o OntoLattes, que foi a construção de uma ontologia, no formato OWL, para comportar os dados dos currículos dos pesquisadores; sobre o SemanticLattes que realiza as tarefas de importação de currículos e lista de veículos de publicações científicas em duas ontologias (descritas inicialmente em DAML+OIL e em seguida OWL), permitindo consultas às instâncias, ele possui um motor de busca que processa a pergunta em linguagem natural e o software, por meio de identificação das palavras-chave, reconhece a pergunta e faz a respectiva consulta em SPARQL. Já o ScriptLattes, que é um software que cria relatórios gerenciais obtidos a partir de um conjunto de

currículos em formato HTML ou XML, não trabalhou com ontologia. As estruturas de dados foram construídas utilizando o conceito de orientação à objetos. Ele foi de relevância no mundo acadêmico e científico, sendo confundido muitas vezes com uma ferramenta que foi desenvolvida e disponibilizada pelo CNPq.

4. E por fim, o projeto Sucupira, que tem por objetivo a extração de informações da Plataforma Lattes para identificação de redes sociais acadêmicas. Uma das principais funcionalidades deste sistema, Sucupira, é o gerenciamento de uma lista de pesquisadores definida pelo usuário, sendo possível visualizar um mapa contendo o endereço profissional dos pesquisadores, um gráfico sumarizado do número de publicações por ano e tipo, e um grafo relacionando os pesquisadores a outros currículos.

Galego (2013) desenvolveu uma ferramenta denominada de Dynamic Lattes que reutiliza as funcionalidades dos trabalhos citados anteriormente e incorpora outras funcionalidades como a possibilidade de alteração do conteúdo dos dados do relatório sem necessidade de alteração da apresentação, a inclusão do relatório de dados inconsistentes, possibilidade de associar uma orientação a formação de algum membro e resumo da comparação dos dados informados pelo orientador com o orientado.

Uma das questões a ser observada nesse levantamento é com relação ao núcleo das pesquisas mencionadas, conceitos de web semântica e ontologia, não foram encontradas pesquisas sobre a anotação semântica, utilizando *Linked Open Data*. Também pode-se destacar a sugestão de futuro trabalho mencionado por esse último autor, (GALEGO, 2013) que é," explorar as funcionalidades de *Linked Data* para que seja possível integração com outras bases de conhecimentos.", que vem ao encontro com o núcleo desta pesquisa (anotação semântica com *Linked Open Data*).

3 Metodologia

Com a fundamentação, foi possível desenvolver a apresentação de conceitos e ferramentas sobre anotação semântica e, em seguida, propor um modelo de um arcabouço e um sistema para a execução da anotação semântica dos documentos do Lattes. O arcabouço representa um modelo da arquitetura necessária para implementar anotações automáticas utilizando *Linked Open Data* através de qualquer técnica. Um experimento comparativo entre as duas ferramentas de extração de entidades identificadas na pesquisa foi necessário para identificar a efetividade do processo de localização e extração dos termos nos documentos do Currículo Lattes.

Quanto ao modelo padrão de *framework* de anotação semântica, estudos (FAFALIOS; PAPADAKOS, 2014; SANTOS NETO, 2009; ARAÚJO

FONTES et al., 2013) indicam um objeto responsável pela análise e extração dos termos de um documento e um objeto responsável pela criação de um documento anotado. A variação encontra-se no quesito da base de dados utilizada para efetuar o mapeamento dos termos encontrados, que no caso de Santos Neto (2009) e Araújo Fontes et al. (2013) utilizam ontologias específicas para o desenvolvimento dos seus trabalhos e no caso do Fafalios e Papadakos (2014) e Zhang et al. (2013) utilizam as ontologias disponíveis no LOD. Quanto a forma de anotar e salvar as anotações, Tao et al. (2013), Saleh e Al-khalifa (2009) e Virgilio et al. (2013) utiliza o formato RDF e salva as anotações de forma não intrusiva e Mendes et al. (2011) utiliza o RDFa também salvando de forma não intrusiva, e além desses, tem o trabalho de Butuc (2009) que utiliza RDF, JSON e microformatos salvando da mesma forma.

Com relação às ferramentas utilizadas nas pesquisas para anotar semanticamente os documentos ou em alguns casos, mencionadas como solução para trabalhar com anotação semântica são AraTation, BioNotate, KIM, Luwak, MnM - SemTag- MUSE, NCBO Annotator, OntoAnnotate, OpenCalais, Rich Snippets, Roseann, Semantator, Semantic Web Annotation Framework, TaToo, Theophrastus, DBpedia Spotlight e GonTogle.

Para trabalhar com extração de entidades (EE), foi identificado nos artigos do referencial teórico os seguintes extratores de entidade: AIDA, Alchemy API, Amilcare, Apache OpenNLP, DBpedia Spotlight, Extrator de Stanford, GATE ANNIE, Lingpipe, Ltasks, LUPEDIA, Mgrep, NERD-ML, OroMatcher, Rembrandt, TextRazor, Whatizit, Wikepedia Miner, YODIE, ZEMANTA.

Na pesquisa realizada por Reeve e Han (2005) os autores compararam ferramentas para anotação semântica disponíveis até o ano da sua pesquisa, ressaltaram suas características e apuraram a eficácia de suas anotações. Os autores destacaram que a ferramenta MnM, que utiliza aprendizagem de máquina na identificação das entidades, como a de melhor desempenho e a de pior a Onto-O-Mat. Em suas conclusões os autores informaram que algoritmos de aprendizagem de máquina são mais efetivos do que os métodos baseados em padrões, porém os sistemas baseados em regras podem possuir uma performance melhor do que os sistemas baseados em aprendizagem de máquina.

Pode-se concluir da fundamentação teórica que para criar um arcabouço de anotação semântico é fundamental um componente de extração de entidade (EE) e um componente que execute a anotação no documento corrente com entidades interligadas às ontologias desejadas. Quanto aos extratores de entidades para a língua portuguesa, o trabalho de Derczynski et al. (2014) enriquece essa pesquisa, citando a solução para esse problema com o EE TextRazor e ou Alchemy que além dos idiomas como o inglês, espanhol, italiano, russo, alemão, também oferecem um serviço de identificação de entidade para o idioma português.

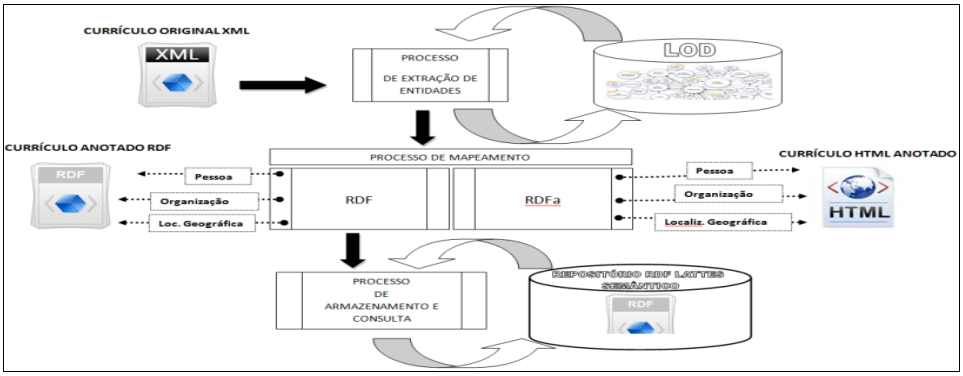
3.1 Arcabouço de anotação semântica automática

A Figura 2 representa a proposta implementada da arquitetura conceitual de anotação semântica e tem como objetivo identificar os componentes, suas relações e etapas que farão parte desta pesquisa.

Podemos explicar o processo nas seguintes etapas:

1. Os dados que alimentam o sistema são documentos/currículos Lattes no formato XML.
2. Processo de Extração tem a função de extrair os termos do currículo, utilizando o componente externo. Ele é uma ferramenta de extração que presta o serviço de extração de entidade via *web*, um *web service*, possui uma abordagem aprendizagem de máquina (NLP), de domínio geral, interligada com as bases abertas do DBPedia e Freebase, porém a licença de uso não é gratuita. Nessas condições, a quantidade de dados a serem utilizados por mês é restrito. O resultado do processo dessa etapa é enviado para o Processo de Mapeamento.

   Figura 2 – Modelo conceitual do projeto



Fonte: Dados da pesquisa.

3. Mapeamento é realizado de maneira que possibilita identificar entidades de um domínio de interesse no currículo. O componente de mapeamento semântico foi desenvolvido utilizando os dados gerados pelo módulo anterior, Processo de Extração, afim de gerar um documento em uma estrutura RDFa e RDF Turtle. Essas estruturas permitem que o documento seja entendido tanto pelas máquinas quanto pelas pessoas e ainda, realização de consultas SPARQL. A geração do documento RDF Turtle, com as triplas de anotações identificadas no Currículo Lattes XML, seguiu o padrão do vocabulário Anotação Aberta[2] (OA).

[2] *Open Annotation Data Model* é uma estrutura para a criação de associações entre recursos relacionados que está de acordo com a arquitetura da *World Wide Web.*

4. No componente Armazenamento e Consulta, é realizado o armazenamento do documento RDF de triplas em um banco de dados (*tripleStore*) que fornece um modo padrão para compartilhamento de dados, intercâmbio, permite consultar e manipular os dados usando a linguagem de consulta SPARQL (TAO et al., 2013).

No fim do processo temos o currículo Lattes anotado com os dados semânticos, legível para as pessoas e máquinas, abrindo possibilidades para os motores de busca inteligentes e a realização de consultas sofisticadas.

3.2. Experimento

Na fundamentação dessa pesquisa encontramos dois Extratores de Entidades para o idioma português: AlchemyApi e TextRazor. São softwares proprietários, mas nessa pesquisa será utilizada a versão gratuita que possui limite de uso (1000 chamadas por dia para o Alchemy e 500 chamadas por dia para o TextRazor). Esse experimento tem como finalidade identificar qual ferramenta reconhecerá mais entidades no Currículo Lattes. É necessário efetuar um teste comparativo entre as ferramentas, identificar qual retorna uma quantidade maior de entidades e informações sobre a entidade extraída.

Quanto à amostra selecionada, a seleção dos documentos Lattes é realizada diretamente no *site* do CNPq, na Plataforma Lattes, selecionando currículo por currículo para realizar o *download* do documento XML, com validação através de *captcha* para cada documento. Atualmente, pela dificuldade que o *captcha* impõe no processo de *download* de documentos na Plataforma e pela própria limitação das ferramentas de extração de entidades (quantidade de strings que podem ser processadas), resolve-se testar aleatoriamente 107 currículos. São documentos que estão no seguinte contexto: de pesquisadores com doutorado e que desempenham atividade de pesquisa e ensino, há descrição no resumo dos documentos, escrito no idioma português e existe número identificador do Currículo Lattes no documento (arquivo XML).

Com o objetivo de definir o tamanho da amostra necessária de currículos que devem ser processados pelos extratores de entidades, para identificar a efetividade deles no processo de extração de entidade, de modo a ter 95,5% de confiança e uma margem de erro máxima de 2%, utilizaremos a fórmula finita de amostragem sobre os documentos que possuímos. O tamanho da amostra necessária está representado pelas fórmulas:

$$n = \frac{\mathrm{N}.Z^2.\mathrm{p}.(1-p)}{(N-1).\mathrm{e}^2 + Z^2.\mathrm{p}.(1-\mathrm{p})}$$

$$n = \frac{(2^2) * 50 * 50 * 107}{5^2 * (107-1) + 2^2 * 50 * 50} = 84{,}585$$

Definida a amostra, para ter-se conhecimento sobre a quantidade de entidades existente em cada documento Lattes é necessário identificar manualmente os termos que estão nesses documentos e existem na base aberta (LOD) do DBpedia português (pt.dbpedia.org). Então, para cada documento da amostragem foi realizada a verificação da existência ou não dos seus termos no mapeamento no DBpedia português, e caso exista, foi anotado o termo, o link do endereço no LOD (pt.dbepedia.org[3]) e o tipo de entidade que está identificada no LOD (entidade:Brasil; link: http://pt.dbpedia.org/resource/Brasil; tipo:thing)[4]. Exclui-se do conjunto encontrado, a classificação do tipo de entidade e termo, que por acaso destaque mais de uma vez dentro do mesmo documento. Essa verificação foi realizada para todos os resumos dos currículos selecionados, totalizando o número de entidades encontradas manualmente por identificador (ID) do Currículo Lattes. Após a coleta manual, utilizando a versão gratuita de ambos os extratores, cria-se um código em linguagem PHP para utilizar os serviços de extração de entidade do *Alchemy* e *TextRazor* de maneira a processar o resumo de cada currículo, retornando os dados do termo extraído e totalizando o número de entidades localizadas.

A Tabela 1 representa a totalização de entidades encontradas em todos os documentos no processo de extração de entidades realizada manualmente, pelo *Alchemy* e *TextRazor* consecutivamente.

Tabela 1 – Experimento: Total de Entidades Extraídas

TOTAL DE ENTIDADES EXTRAÍDAS		
MANUAL	**ALCHEMY**	**TEXTRAZOR**
7406	1005	5537

Fonte: Dados da pesquisa.

Após a totalização de entidades extraídas manualmente, pelo Alchemy e pelo TextRazor, pode-se também, contrastar a quantidade encontrada entre os extratores de entidades, a quantidade de entidades que os extratores automáticos conseguiram identificar e que também foram identificados manualmente, isso é, confrontar os termos encontrados entre os três processos de extração executados.

A interseção de entidades extraídas entre a anotação manual e o Alchemy é de 455 entidades. A quantidade encontrada entre a extração manual e o TextRazor é de 4.090 e entre o Alchemy e o TextRazor de 536 entidades.

3.3 Resultados do experimento

Utilizando como referência o trabalho de Reeve e Han (2005) que utiliza a medida padrão de *Precision, Recall*, e *F-measure* na determinação da eficácia das Plataformas de Anotação Semântica (PAS). Aproveitam-se

[3] Acesso em: 30 jan. 2016.
[4] Acesso em: 30 jan. 2016.

os dados obtidos nestes testes empíricos para determinar a eficácia das extrações de entidades calculando a média padrão para as entidades extraídas pelo *Alchemy e TextRazor:*

a) A fórmula que representa a Precisão, fração de instâncias recuperadas que são relevantes:

$$\textbf{Precisão} = \frac{Intersecao\ da\ Extracao\ Manual\ x\ Extracao\ Automatica}{Número\ de\ Entidades\ da\ Extracao\ Automatica}$$

b) A que representa a Revocação, fração de instâncias pertinentes que são recuperadas:

$$\textbf{Revocacão} = \frac{Intersecao\ da\ Extracao\ Manual\ x\ Extracao\ Automatica}{Número\ de\ Entidades\ da\ Extracao\ Manual}$$

c) E a fórmula que representa o F-measure, a média harmônica de precisão e revocação:

$$\boldsymbol{F-measure} = 2 * (\frac{Precisao}{Revocacao})$$

A Tabela 2 representa os valores de precisão, revocação e média harmônica que foram obtidos com o somatório de entidades localizadas pelas ferramentas extração de entidade Alchemy e TextRazor. É possível observar pelos valores apresentados que o TextRazor possui mais resultados relevantes quanto a extração de entidade que o AlchemyApi.

Tabela 2 – Experimento: Valores de Precisão, Revocação e Média Harmônica dos Extratores de Entidade Alchemy e TextRazor

API EXTRATOR	*Precisão*	*Recall*	*F-measure*
EE AlchemyAPI	*0,453*	*0,061*	*0,246*
EE TextRazor	*0,739*	*0,552*	*2,209*

Fonte: Dados da pesquisa.

O AlchemyApi, após a identificação de uma entidade, retorna o termo extraído e a classificação do tipo de entidade desambiguada. O TextRazor retorna para cada entidade extraída, a classificação do tipo de entidade desambiguada e o *link* onde está mapeado o termo no LOD. O TextRazor extraí as entidades e as classifica utilizando o tipo mais adequado ao contexto. O TextRazor conseguiu identificar e extrair as datas (ano em formato de quatro dígitos), classificando-as como do Tipo *Date* e também os números, classificando-os como do Tipo *Number*, porém nesse tipo de entidade não há retorno de informações como endereço no LOD.

Como já mencionado na fundamentação, o processo manual de extração de entidade, desambiguação e classificação do tipo de entidade é custoso. Os Extratores de Entidades surgem com o objetivo de agilizar esse processo. No sentido de identificar qual extrator é mais eficiente, o teste de aferição entre os extratores de entidades AlchemyApi e TextRazor, com textos no idioma português, apontou o TextRazor como o

mais eficiente para extrair entidades nos documentos do Lattes, porque perante as entidades extraídas nos documentos o TextRazor tem uma efetividade maior em relação ao Alchemy. Além disso, sempre retorna o tipo da entidade desambiguada e seu *link* no endereço LOD (DBpedia e outras bases como Freebase). Com esse experimento, a ferramenta de extração automática de entidade, TextRazor, torna-se essencial e qualificada para agilizar e atuar no processo de anotação automática de entidades dos Currículos Lattes.

4 Implementação do sistema

Com as definições da ferramenta de extração e do modelo, foi criado o sistema de anotação semântica dos currículos da Plataforma Lattes, denominado de Lattes Web Semântico[5].

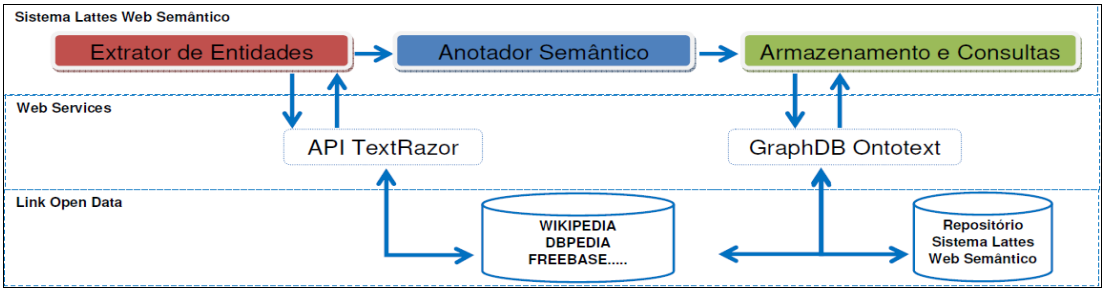
A Figura 3 representa a visão geral dos componentes do sistema. Esse sistema executa a funcionalidade de mapeamento semântico com a geração dos arquivos HTML (RDFa) e RDF com as entidades localizadas e anotadas baseadas no LOD.

Figura 3 – Visão Geral dos Componentes do Sistema Lattes Web Semântico

Fonte: Dados da pesquisa.

O Sistema Lattes Web Semântico (LattesWS) foi desenvolvido na linguagem PHP, auxiliado por um *Web Service* (TextRazor) para a extração de entidade e um gerenciador de triplas RDF. A Figura 4 representa uma visão geral da iteração entre os componentes que integram esse Sistema.

Figura 4 – Visão Geral da Iteração entre os componentes do Sistema Lattes Web Semântico



Fonte: Dados da pesquisa.

[5] Disponível em: <www.wssistemas.com.br/latteswss/>. Acesso em: 30 jan. 2016.

O sistema é composto pelas seguintes unidades:

1. Extrator de Entidades: Esse componente tem como objetivo identificar as entidades, seus tipos e links dos bancos de dados abertos (LOD) em um texto.

2. Anotador Semântico (Anotador Semântico RDFa e Anotador Semântico RDF): Esse componente tem como objetivo receber os dados gerados pelo Extrator de Entidades, manipulá-los e gerar um arquivo RDFa e RDF Turtle.

3. Armazenamento das Triplas e Consulta Semântica: Foi definido para armazenar e gerenciar os arquivos RDF Turtle um banco de dados que armazena triplas RDF e possibilita efetuar alterações e consulta semânticas sobre os dados armazenados.

A efetividade do sistema acontece a partir da existência da anotação dos arquivos gerados no formato RDFa, nas consultas geradas sobre os dados RDF Turtle armazenados e na interligação dos dados com LOD. Como prova do conceito, o sistema disponibiliza os arquivos anotados e consultas semânticas. Os arquivos semânticos são armazenados no diretório da aplicação e podem ser observados no próprio *site*.

5 Considerações finais

A implantação da Web Semântica possibilita aumentar o significado de um conteúdo, tornando-o interpretável por humanos e aplicações. Como consequência, viabiliza um maior entendimento da estrutura do documento e otimiza a recuperação de informações. A anotação semântica é a forma pela qual podemos implantar uma Web semanticamente compreensível por pessoas e computadores.

Com o experimento empírico foi possível identificar qual extrator é o mais apropriado ou efetivo no processo de identificar, extrair, desambiguar e classificar a entidade dentro do contexto da língua portuguesa. O EE TextRazor mostrou-se mais eficiente uma vez que seus valores de precisão, revocação e média harmônica foram maiores do que o do Alchemy. Outra característica marcante no TextRazor, além dessa eficiência na identificação das entidades, são as informações sobre a entidade na base aberta (denominada de *Linked Open Data*). Essas informações são relevantes para essa pesquisa porque na criação da PAS utilizando LOD é necessário ter conhecimento do *link* (endereço) onde está mapeado a entidade, pois só com essa informação é possível anotar a entidade utilizando o LOD.

Nessa pesquisa, foi desenvolvido um Sistema, Lattes Web Semântico[6], que executa anotação semântica automática de entidades para os documentos da Plataforma do Currículo Lattes do CNPq utilizando dados de bases abertas. O sistema é composto por um componente de

[6] Disponível em: <www.wssistemas.com.br/latteswss/>. Acesso em: 30 jan. 2016.

extração de entidade, um anotador semântico e um componente de armazenamento e consulta de dados. Possibilita a disponibilização e indexação dos Currículos Lattes na Web através dos documentos HTML com RDFa embutido, disponibiliza as informações em RDF Turtle para serem utilizados em outros sistemas e domínios e, por fim, expande a possibilidade de buscadores semânticos atuarem sobre esses Currículos.

Criar um sistema de anotação semântica de entidades não é uma tarefa trivial. É obrigatório, devido à quantidade de documentos que há no cenário Lattes, a existência de ferramentas que trabalhem de forma automática (SANTOS NETO, 2009). Na análise da literatura, foi obtida a informação do Extrator de Entidade, TextRazor, que possibilita encontrar entidades em textos da língua portuguesa de forma automática e a sua eficiência foi testada no experimento empírico. Nessa pesquisa encontraram-se as informações necessárias para a compreensão do cenário das bases abertas (LOD), seu funcionamento, acesso e particularidades; entendimento de uma outra estrutura de armazenamento de dados e consulta por meio de banco de dados de triplas utilizando a linguagem SPARQL.

Com os estudos teóricos, desenvolvimento das etapas da concepção do sistema de anotação e observando os objetivos de cada componente, assim como apresentado por Reeve e Han (2005), conclui-se que para uma plataforma de anotação semântica (PAS) o componente de extração de entidade (EE) é um objeto de grande relevância para o sucesso das anotações.

Atentar que para o futuro das PAS, que terão como papel oferecer suporte, dados, para as máquinas de busca arquitetadas como buscadores semânticos, pode-se dizer que anotar corretamente é disponibilizar informações integras, completas e corretas sobre um termo. Então, mais importante do que velocidade para encontrar entidades durante uma anotação automática de texto é a certeza de que o termo/palavra está caracterizado na entidade correta.

Diante das limitações desse desta pesquisa, pode-se recomendar os seguintes trabalhos futuros: Criação de um extrator de entidade específico para a língua portuguesa; Utilização da junção de outros extratores de entidades para otimizar a quantidade de identificação dos termos/palavras nos LOD's; Criar interface de operação de anotação manual para o usuário, possibilitando aumentar o número de termos anotados; Continuar a extração das outras partes do Lattes, disponibilizando-as para a anotação semântica; Verificar a possibilidade de interligar com os dados do projeto do Portal Brasileiro de Dados Abertos e Utilizar agentes de pesquisa inteligente (buscador semântico) sob os documentos do lattes semântico.

O arcabouço desenvolvido nessa pesquisa juntamente com o experimento do extrator de entidade demonstrou a aplicação prática dos conceitos relacionados a Anotação Semântica Automática de Entidades utilizando *linked open data* (dados abertos conectados). O sistema desenvolvido nessa pesquisa atendeu o objetivo do LOD em identificar

conjuntos de dados disponíveis sob licenças abertas, no caso os documentos Lattes, e convertê-los para RDF de acordo com as diretrizes do Linked Data (CHRISTIAN BIZER, 2009).

Referências

ARAÚJO FONTES, C. A.; CAVALCANTI, M. C.; MOURA, A. M. D. C. An ontology-based reasoning approach for document annotation. In: SEMANTIC COMPUTING (ICSC), 7., 2013; INTERNATIONAL CONFERENCE ON IEEE, 7., 2013. *Proceedings...* Irvine; California, USA: IEEE, 2013. p. 160-167.

BELLOZE, K. T. et al. An evaluation of annotation tools for biomedical texts. *ONTOBRAS-MOST*, v.?, n.?, p. 108-119, 2012.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, v. 284, n. 5, p. 34-43, 2001.

BONIFACIO, A. S. *Ontologias e consulta semântica*: uma aplicação ao caso lattes. 2002. 85f. Dissertação (Mestrado em Ciência da Computação) - UFRGS, Porto Alegre. (2002).

BUTUC, M.-G. *Semantically enriching content using op*encalais. Suceava: USV, 2009.

CASTANO, A. C. *Populando ontologias através de informações em html*: o caso do currículo Lattes. 2008. 100f. Mestrado (Ciência da Computação) - Universidade de São Paulo, São Paulo, 2008.

CEWEB. *Centro web brasil*. 2016. Disponível em: <http://ceweb.br/>. Acesso em: 30 jan. 2016.

CHRISTIAN, B.; HEATH, T.; BERNERS-LEE, T. Linked data the story so far. *Int. J. Semantic Web Inf. Syst.*, v. 5, n. 3, p. 1-22, 2009.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPQ). *Plataforma Lattes Cnpq*. 2014. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 16 mar. 2014.

DERCZYNSKI, L. et al. Analysis of named entity recognition and linking for tweets. *Information Processing and Management,* v. 51, n. 17, p. 32-49, 2014.

ELLER, M. P. *Anotação semânticas de fontes de dados heterogêneas*: um estudo de caso com a ferramenta Smore. 2014. Dissertação (Mestrado em Computação) - Universidade Federal de Santa Catarian, Departamento de Informática e Estatástica, 2014.

FAFALIOS, P.; PAPADAKOS, P. Theophrastus: on demand and real-time automatic annotation and exploration of (web) documents using open linked data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, v. 29, p. 31-38, 2014.

FONTES, C. A. et al. *Recuperação de informações em documentos anotados semanticamente na área de gestão ambiental*. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL – ONTOBRAS, 3., 2010. *Anais...* Florianópolis: UFSC, 2010. p. 43–52.

GALEGO, E. F. *Extração e consulta de informações do currículo lattes baseada em ontologias*. 2013. 70f. Dissertação (Mestrado em Ciência da Computação) - Universidade de São Paulo, São Paulo, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-18122013-080644/pt-br.php>. Acesso em: 20 dez. 2014.

GOV, W. D. A. *Dados abertos governamentais.* 2014. Disponível em: <http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf>. Acesso em: 16 mar. 2016.

MENDES, P. N. et al. DBpedia spotlight: shedding light on the web of documents. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 7., Graz, 2011. *Proceedings...* Graz: ACM, 2011. p. 1-8.

OREN, E. et al. What are semantic annotations. [s.l.]: Citeseer, 2006

PLANETDATA. *Linked open data cloud diagram*. 2014. Disponível em:

<http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/>. Acesso em: 16 mar. 2016.

PRIMER, W. R. *Rdfa 1.1 primer*. 2. ed. 2014. Disponível em: <http://www.w3.org/TR/2013/NOTE-rdfa-primer-20130822/>. Acesso em: 16 mar. 2016.

REEVE, L.; HAN, H. *Survey of semantic annotation platforms*. Santa Fe: ACM, 2005. p. 1634-1638.

SALEH, L. M. B.; AL-KHALIFA, H. S. AraTation: an Arabic semantic annotation tool. In: INTERNATIONAL CONFERENCE ON INFORMATION INTEGRATION AND WEB-BASED APPLICATIONS & SERVICES, 11., *Proceedings…* Malaysia: Kuala Lumpur, 2009. p. 447-451.

SANTOS NETO, G. M. dos. *Anotação semântica de recursos web baseada em ont*ologias. 2009. Mestrado (Dissertação em Ciência da Computação) - UFAM, Instituto de Ciências Exatas, Programa De Pós-Graduação Em Informática, 2009.

TAO, C. et al. Semantator: semantic annotator for converting biomedical text to linked data. *Journal of Biomedical Informatics*, v. 46, n. 5, p. 882-893, 2013.

VIRGILIO, R. D. et al. A reverse engineering approach for automatic annotation of web pages. *Multimedia Tools and Applications*, v. 64, n. 1, p. 119-140, 2013.

WORLD WIDE WEB CONSORTIUM (W3C). *W3C-Schema1.1:* Rdf schema 1.1. 2014. Disponível em: <http://www.w3.org/TR/2014/REC-rdf-schema-20140225/>. Acesso em: 16 set. 2015.

ZHANG, Z.; CHEN, S.; FENG, Z.. Semantic annotation for web services based on DBpedia. In: INTERNATIONAL SYMPOSIUM ON SERVICE-ORIENTED SYSTEM ENGINEERING, 7., 2013. *Proceedings...* Redwood City: IEEE, 2013. p. 280-285.