DOI: 10.5748/9788599693131-14CONTECSI/PS-4762

MODEL OF DATA EXTRACTION IN THE INNOVATION ENVIRONMENTS OF THE STATE OF SÃO PAULO BASED ON SEMANTIC TECHNOLOGIES

Melissa Cordeiro Cavalcanti (Centro Universitário Eurípides de Marília, São Paulo, Brasil) - melcavalcanti@gmail.com
Fábio Dacêncio Pereira (Centro Universitário Eurípides de Marília, São Paulo, Brasil) - prof.fabiopereira@gmail.com
Elvis Fusco (Centro Universitário Eurípides de Marília, São Paulo, Brasil) - elvisfusco@univem.edu.br
Marcos Luis Mucheroni (Universidade de São Paulo, São Paulo, Brasil) - mucheroni.marcosl@gmail.com

The digital information environments of the innovation actors of the State of São Paulo offer a complex set of heterogeneous information and without a pattern of representation and retrieval. The relationship between government, companies and universities is complex because they have different points of view and the information found in this medium can be scattered and without a specific format. In view of this scenario, an Intelligent Extraction Agent was implemented through a Computational Robot capable of extracting semistructured information from Web pages in the context of the innovation environments of the State of São Paulo, built with the help of the Jsoup library and the Jena Framework. This extracted information is inserted into the predefined RDF structure, and queries can be performed according to the syntax of the words and for a time interval using the SPARQL language. The purpose of this implementation was to propose a semantic model of extraction and schematization of the semistructured data that are extracted and consulted by the user, so that the retrieval based on Semantic Web concepts of the information that adds value to the State Innovation Actors from Sao Paulo. The built-in Robot Extractor has as informational source data taken from the CNPq website and demonstrates the direction of a way to improve the extraction and schematization of semi-structured and unstructured data. According to the results obtained, it can be observed that, for future works, a search performed by the user can be objectified the moment a specific ontology is inserted and the data fusion, to build the interrelationship between the information related to the Actors Of Innovation, taken from the scenario that is adopted as an informational source.

Keywords: Information Extraction Agents; Innovation; Semantic Web.

Introdução

O Governo do Estado de São Paulo busca incentivar o cenário de inovação com programas e leis que foram criados e estabelecidos com a intenção de melhorar o relacionamento entre governo, empresas e universidades, já que para a cultura de inovação se desenvolver é necessário um bom envolvimento entre estes atores de inovação.

As leis e acordos criados e estabelecidos pelo governo estadual, a partir do início anos 2000, têm o objetivo de incentivar e melhorar a relação e a integração entre empresas, instituições governamentais e educacionais com o intuito de ampliar as informações e os produtos existentes que são relacionados ao conceito de inovação tecnológica. Assim, são capazes de estabelecer procedimentos entre esses agentes, que apoiam projetos nos ambientes de desenvolvimento de inovação e tecnologia.

Os parques tecnológicos, as incubadoras de empresas de base tecnológica e os centros de inovação tecnológica são empreendimentos que procuram incentivar a relação entre os atores de inovação por meio de incentivos. Estes são capazes de disponibilizar suportes, de diversas formas, para que haja o desenvolvimento de projetos embasados nos conhecimentos tecnológicos e inovação e apoiando o desenvolvimento de empresas de tecnologia a curto e longo prazo. O núcleo de inovação tecnológica tem como finalidade gerir a política de inovação das instituições científicas e tecnológicas do estado de São Paulo.

Os robôs de busca e extração de dados são agentes computacionais capazes de percorrer um determinado cenário Web que é restringido a ele em sua implementação, procurando e decidindo as informações que possuem maior grau de relevância e serão extraídas de acordo com os padrões que são estabelecidos em seu desenvolvimento. Os agentes de extração podem realizar as atividades destinadas a eles manipulando as informações retiradas do meio que relacionam localidade, período de tempo e palavras, por exemplo.

O fato da Web ter crescido e se desenvolvido de maneira não centralizada, mostra que as informações existentes neste meio necessitam de técnicas que são capazes de classificar, organizar e estruturar os dados que são buscados e extraídos deste ambiente. Segundo Gruber (apud Oliveira e Werneck, 2003), a ontologia é capaz de definir o contexto e os vocabulários em um domínio com múltiplos agentes e podem servir como base para comunicação entre eles na extração de informações.

Com o intuito de traçar um modelo semântico para maximizar a extração e esquematização dos dados semiestruturados que são recuperados pelo usuário, para que futuramente seja realizada a utilização das informações que agregam valor aos Atores de Inovação do Estado de São Paulo, foi implementado um Robô Extrator capaz de extrair informações semiestruturadas de páginas Web de Chamadas Públicas Abertas do site do CNPq, com o auxílio da biblioteca Jsoup e do Framework Jena. Essas informações extraídas da Web são inseridas na estrutura RDF pré-definida e podem ser realizadas consultas de acordo com a sintaxe das palavras e por um intervalo de tempo utilizando a linguagem SPARQL.

A extração das informações semiestruturadas de uma página web com auxílio do Jsoup demonstrou que esta biblioteca disponível para Java é capaz de contornar essa limitação por não existir um padrão de formato pré-definido para os dados que são encontrados. Assim, como a utilização do Jena foi capaz de mostrar que esses dados extraídos podem ser persistidos na estrutura RDF construída e manipulados de acordo com o que tiver relevância para o usuário, com o uso de aplicações que podem ser disponibilizadas pelo próprio *framework*, como é o caso do SPARQL que foi usado para realizar as consultas de acordo com a sintaxe das palavras e por um período de tempo.

1. Cenário de Inovação Tecnológica no Estado de São Paulo

Nesta seção é apresentada a ambientação do cenário de inovação tecnológica no estado de São Paulo. O objetivo é mostrar alguns ambientes e projetos que têm como propósito auxiliar o relacionamento entre os agentes de inovação, que também fazem parte de todo processo para a recuperação de informações.

1.1 Ambiente de Inovação Paulista

O campo de desenvolvimento tecnológico no Brasil vem evoluindo pela modificação das relações existentes entre pesquisadores, indústrias e o meio acadêmico, estes conhecidos como os atores de inovação. A partir do início dos anos 2000, os governos federais e estaduais brasileiros começaram a desenvolver maneiras para que a evolução da ciência, pesquisa e desenvolvimento tecnológico ocorresse de forma mais eficaz, e com a existência do relacionamento entre os atores.

Políticas públicas, programas e órgãos governamentais, em esfera nacional, estão sendo criados e reestruturados com a finalidade de unir e articular os as iniciativas privadas, universidades e as diferentes esferas existentes no governo, seja em questão nacional, estatal ou municipal. Pereira (2016) destaca como alguns desses programas:

- Marco da Ciência Tecnologia e Inovação: regulamenta a Emenda Constitucional 85, definindo parcerias de longo prazo entre os setores público e privado, dá maior flexibilidade de atuação às Instituições Científicas, Tecnológicas e de Inovação (ICTs) e respectivas entidades de apoio.
- Programa Nacional de Apoio às Incubadoras de Empresas e aos Parques Tecnológicos (PNI) do Ministério da Ciência, Tecnologia e Inovação(MCTI).
- Financiadora de Estudos e Projetos (FINEP), empresa pública criada em julho de 1967 e vinculada ao MCTI. Tem como missão "Promover o desenvolvimento econômico e social do Brasil por meio do fomento público à Ciência, Tecnologia e Inovação em empresas, universidades, institutos tecnológicos e outras instituições públicas ou privadas."
- Política Industrial, Tecnológica e de Comércio Exterior – PITCE - A PITCE379 foi instituída com o objetivo de aumentar a competitividade das empresas brasileiras, mediante elevação dos níveis de eficiência e produtividade, fomento à capacidade inovadora e estímulo às exportações.
- Lei de Inovação e Lei do Bem: (a) incentiva parcerias em P&D entre universidades, instituições de pesquisa e empresas; (b) regula a transferência de tecnologia e a criação de incubadoras;(c) permite compartilhar equipamentos, infraestrutura e pessoal em atividades de desenvolvimento de novas tecnologias; e (d) estabelece subsídios e recursos para tais atividades.
- Conselho Nacional e Agência Brasileira de Desenvolvimento Industrial: Para aprimorar a coordenação institucional e, principalmente, incentivar a inovação e os gastos das empresas privadas em P&D, foram estabelecidos dois novos órgãos: o Conselho Nacional de Desenvolvimento Industrial (CNDI) e a Agência Brasileira de Desenvolvimento Industrial (ABDI).

Segundo Bruno Rondani[1] (apud Mendonça, 2012: 70) de acordo com a visão instituída até aquele momento "a ciência só é boa se for para gerar conhecimento. [. . .] O

[1] Diretor do Centro de Pesquisa e Inovação Sueco-Brasileiro (Cisb) em 2012

ideal é que haja todo um sistema de financiamento e cooperação entre os atores responsáveis pela inovação tecnológica [. . .]". Assim, o governo do estado de São Paulo desenvolveu e instalou programas e legislações como forma de incentivo ao relacionamento acadêmico, industrial e governamental, com a intenção de melhorar o desenvolvimento no campo de inovação estadual.

De acordo com a Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação do estado de São Paulo, o Centro de Inovação Tecnológica (CIT) é um local criado com o intuito de realizar o estímulo ao crescimento e à competitividade das micro e pequenas empresas com auxílio do avanço tecnológico que promove a interação entre empreendedores e pesquisadores; é capaz de oferecer mecanismos e serviços de suporte ao processo de inovação das empresas e teve como apoio à sua instalação o lançamento da Rede Paulista de Centros de Inovação Tecnológica (RPCITec).

Os objetivos da RPCITec são realizar ações estimulantes à cultura de inovação e aos centros de inovação tecnológica integrante dela a realização de pesquisa, desenvolvimento e engenharia de produtos e/ou processos; facilitar e estimular a consolidação de parcerias entre esses centros de inovação tecnológica com empresas e organizações que participam da área produtiva; e realizar todo o apoio necessário para o desenvolvimento, seja por meio de capacitações, treinamentos e eventos, como disponibilização de serviços tecnológicos.

A Rede Paulista de Incubadoras de Empresas de Base Tecnológica (RPITec), instituída pelo decreto Nº 56.424 de 23 de Novembro de 2010, foi criada com a intenção de apoiar, fortalecer e estimular a instalação de empresas inovadoras em desenvolvimento de produtos e/ou processos no Estado, já que incubadoras de empresas de base tecnológica, conhecidas como EBTs, são capazes de oferecer espaço físico por determinado período de tempo para empresas da área tecnológica que estão iniciando, oferecendo suporte gerencial e tecnológico; o que gera a interação entre essas empresas e, consequentemente, realiza a troca de informações e a difusão do conhecimento.

Em 25 de março de 2014 foi instituído o decreto Nº 60.286 que regulamenta e institui o Sistema Paulista de Ambientes de Inovação (SPAI). Este, por sua vez, abrange o Sistema Paulista de Parques Tecnológicos (SPTec), a RPITec, a RPCITec e a Rede Paulista de Núcleos de Inovação Tecnológica (RPNIT). Este decreto considera que parques tecnológicos, incubadora de empresas de base tecnológica, centro de inovação tecnológica e núcleo de inovação tecnológica devem apoiar projetos e/ou processos tecnológicos nos ambientes de inovação e tecnologia, incentivando a relação entre os atores de inovação.

O Sistema Paulista de Parques Tecnológicos tem como objetivo estimular o surgimento, desenvolvimento, competitividade e aumento da produtividade de empresas com atividades baseadas no conhecimento, na tecnologia e na inovação, incentivando a interação entre instituições de pesquisa, meio acadêmico, capital de oportunidade e investidores; realizar o desenvolvimento de São Paulo atraindo investimentos para as atividades baseadas no conhecimento e na inovação tecnológica.

1.1.1 Parques Tecnológicos Credenciados Definitivamente

Os Parques Tecnológicos que participam do SPAI recebem inicialmente o credenciamento temporário, o qual permite a atuação destes no cenário tecnológico até que seja aprovado, pela organização regulamentadora, o credenciamento definitivo.

O Parque Tecnológico de São José dos Campos possui Centros Empresariais que abrigam aproximadamente 60 empresas e oferece para essas empresas espaço físico e infraestrutura básica capazes de abrigar suas instalações e seu pessoal. Criado em 2009 por

iniciativa da Prefeitura de São José dos Campos, foi o primeiro parque a ser credenciado definitivamente pelo SPTec no ano de 2010.

Localizado em São Carlos, o ParqTec tem como finalidade a promoção do desenvolvimento regional com otimização do custo da transação realizada por inovação tecnológica e mercado. Contribui de maneira significativa na construção de uma Região de Inovação constituída por universidades públicas e privadas, centros de pesquisas, órgãos de governo e por um conjunto de mais de 180 EBT's.

Responsável por atrair e reter empresas tecnológicas, com destaque para os setores de Saúde, Biotecnologia, Tecnologia da Informação e Bioenergia, o Parque Tecnológico localizado em Ribeirão Preto surgiu de uma parceria entre USP, Prefeitura Municipal e Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação do Estado de São Paulo.

Inserido no SPTec, o Parque Tecnológico de Piracicaba surgiu das diferentes visões de pessoas determinadas nos governos Estaduais e Municipais. Seus objetivos são promover a informação tecnológica, estimular que centros de pesquisa, universidades e empresas cooperem entre si e dar suporte ao desenvolvimento de atividades empresariais.

O Parque Tecnológico de Sorocaba (PTS) criado para atrair e acomodar empresas com base tecnológicas, instituições de ensino e pesquisa, assim como empresas de consultoria ou organizações, públicas e/ou privadas, que possam oferecer serviços de apoio técnico e de mercado, com o intuito de facilitar acesso ao conhecimento e ao mercado, pela aproximação com possíveis desenvolvimentos e inovação tecnológica.

Situado na cidade de Santos, o FPTS realiza a promoção da ciência e tecnologia e age aproximando os centros de conhecimento e o setor produtivo, oferece oportunidade para que as empresas do Estado possam transformar pesquisa em produto. Tem o intuito propagar a cultura da inovação e empreendedorismo com a finalidade de realizar o desenvolvimento sustentável na cidade e na região metropolitana da Baixada.

1.1.2 Centros de Inovação

Os Centros de Inovação possuem como alguns objetivos a promoção da competitividade local e regional, sediar incubadoras de empresas de base tecnológica e laboratórios específicos de acordo com a demanda regional.

Localizado na cidade de Jundiaí, o CITJUN, tem como propósito primário facilitar o desenvolvimento tecnológico e atua agregando e incentivando as ações governamentais, acadêmicas e empresariais da região. Desenvolvido pela Prefeitura de Jundiaí, governo de São Paulo, pelo Sincomércio que é a gestora do Centro de Inovação e pela Companhia de Informática de Jundiaí.

O Centro de Inovação Tecnológica de Marília (CITec-Marília) foi credenciado pelo Governo do Estado em dezembro de 2015 na RPCITec. Esta tem como finalidade promover o fortalecimento e estimular os processos locais e regionais em benefício do desenvolvimento e da competitividade das empresas da região, também oferece um espaço adequado para a pesquisa, desenvolvimento e inovação (P&D&I) de empresas que tenham perfil inovador.

Portanto, percebe-se que as ações realizadas pelo governo estadual com o intuito de melhorar o seu cenário de inovação, tiveram um retorno positivo e facilitou o relacionamento entre os atores de inovação. Entretanto, mesmo com os diversos tipos de incentivos, pelo fato de empresas, centros acadêmicos e agentes governamentais possuírem objetivos muito distintos, a existência da relação entre eles ainda é de grande dificuldade. Assim, faz necessária a continuidade desses projetos e ações, com a finalidade de melhorar a interação entre os atores e o desenvolvimento da inovação tecnológica no estado.

2. Conceitos e Tecnologias Semânticas

Para realização da Extração de Dados nos ambientes Web, por meio de um robô de extração de dados, são utilizados softwares, aplicações e codificações específicas para extração de informações. A seguir apresentam-se informações sobre o cenário, métodos para a estruturação das informações acessadas e biblioteca existente para a busca de dados encontrados em páginas Web.

2.1 Web Semântica

O ambiente Web se desenvolveu de maneira difusa, tendo como prioridade inicial para seu desenvolvimento a construção da rede, fazer com que esta fosse ser acessível e capaz de ser comercializada. Por conta da descentralização gerada durante toda a sua construção e evolução, surge a necessidade de encontrar e interpretar conteúdos específicos para a recuperação das informações.

No resultado de uma busca realizada por um usuário pode estar contido inúmeras informações, sendo estas, relevantes ou não para ele, por conta da vasta quantidade de informações que estão disponíveis neste meio. Esse acontecimento permite que o próprio usuário tenha o poder de decidir e verificar as informações resultantes que tem importância real para serem usadas por ele.

> O principal propósito da Web Semântica é de atribuir significado ou sentido a qualquer conteúdo publicado na internet, através da utilização de metadados, de maneira a tornar as informações na Web interpretáveis por computadores, alcançando assim resultados mais rápidos, inteligentes, eficientes e precisos no compartilhamento de informações. (Prybecs, Gonçalves Junior, Mendes, 2013:6)

Assim, percebe-se que a Web Semântica aparece como um auxílio para que o desenvolvimento de aplicações que sejam capazes de dar como resposta ao usuário informações realmente relevantes para sua busca e que melhor se relacionam com o que é explicitado, e não toda a vasta quantidade de dados que são encontrados de maneira descentralizada que podem não ter valor associado ao que foi especificado.

Segundo Bräscher (2007), a Web Semântica é uma plataforma acessível e universal capaz de permitir que os dados possam ser compartilhados e processados tanto por ferramentas automáticas como por pessoas, são agentes capazes de realizar busca, filtragem e preparação de dados encontrados para o usuário.

De acordo com Dias e Santos (2003), a proposta da Web Semântica não é uma separação da Web atual, mas sim uma extensão dela, baseada em ontologias capazes de descrever os relacionamentos entre os objetos e conter suas informações semânticas para acontecer automatização do processamento pelas máquinas, que não acontecem no momento.

Coneglian (2014) fala que uma das maneiras que a Web Semântica pode ser realizada é realizando uma divisão por camadas para que ela seja aplicada, sendo elas:

- URI (*Uniform Resource Identifier* – Identificador de Recursos Uniforme): conjunto de caracteres para a identificação de um recurso (W3C, 2014b);
- Unicode: define um conjunto e padrão universal de codificação (UNICODE, 2008);

- XML (*Extensible Markup Language* – Linguagem de Marcação Extensível): é um sistema de representação de informação estruturada (W3C, 2014c);
- *Namespace*: um conjunto de nomes, identificada por uma referência URI.
- XML *Schema*: expressam os vocabulários compartilhados e permitem que as máquinas vejam as regras feitas pelas pessoas (W3C, 2014d);
- RDF M&S: um modelo para intercâmbio de dados na web, e tem características que facilitam a fusão de dados (W3C, 2014e);
- RDF *Schema*: um vocabulário para fazer a modelagem de dados de RDF (W3C, 2014f);
- *Ontology*: será tratado com mais clareza ainda neste capítulo;
- *Rules*: nela é feita a conversão das informações que estão dentro de um documento para outro, criando regras de inferência (PRADO, 2004).
- *Logic*: tem a intenção de transformar o documento em uma linguagem lógica, fazendo inferências e funções, para que duas aplicações de RDF sejam conectadas.
- *Proof*: pode-se depois de passar por várias camadas, fazer uma prova deste documento, ou seja, pode-se provar hipóteses a partir das informações.
- *Sig*: assinatura, para verificar a autonomia do documento.
- *Trust*: tendo a assinatura do documento, pode-se saber a confiança nesta informação.

É observado que para realizar buscas mais eficazes e capazes de conseguir resultados com grande valor de relevância, podem ser utilizadas ontologias que organizam os dados da Web de forma que a base de dados seja persistida com informações sobre a relação feita entre os dados extraídos de Fontes Informacionais que não se relacionam, baseando-se na interpretação das respostas realizada por um aplicativo que é usado para tal propósito.

2.2 Ontologias

O conceito de ontologias não existe apenas no campo de Tecnologia da Informação, na realidade, um dos cenários pioneiros na classificação de ontologia foi a Filosofia, que usou desse termo para explicar o ser. Baseado nos apontamentos do curso de Formação de Gestores do Conhecimento da UFBA (2007), Platão foi responsável pelo primeiro modelo de representação do conhecimento baseado nas questões que eram conhecidas até aquele momento sobre os seres vivos.

Para a Computação, o interesse na ontologia teve o intuito de garantir o conhecimento sobre informações relacionadas ao cenário que é usado pelo mecanismo de representação escolhido por um usuário. É observado que a Inteligência Artificial usa desse conceito para realizar descrição de domínios conhecidos e pode servir de auxílio para reuso e compartilhamento das informações que são utilizadas tanto por usuários como por máquinas.

De acordo com Gruber (apud Oliveira e Werneck, 2003: 2) “ontologia é uma especificação explícita de uma conceituação. A conceituação é a organização do conhecimento em forma de entidades e a especificação é a representação dessa conceituação em uma forma concreta”.

Ontologias podem ser classificadas e utilizadas de diversas maneiras, tanto como por nível de generalização como por categorias ou tipos de utilização, dependendo então de como a informação deve ser observada.

A classificação de ontologias desenvolvida por Guarino foi construída de acordo com o nível de generalização, podendo elas serem classificadas como: genérica que descrevem dados gerais como objetos, funções, eventos, tempo, etc; tarefas que descreve um vocabulário de termos que tem relação com atividades, independente do domínio relacionado; domínio é capaz de especificar um vocábulo pertencente ao domínio desejado; aplicação descreve informações necessárias para uma aplicação que dependem tanto de um domínio em específico quanto de uma atividade pertencente a este domínio. (UFBA, 2007)

Outra classificação que pode ser encontrada foi criada por Uschold, construída tendo como base o tipo de conhecimento pode ser dividida em ontologias de representação, domínio e tarefas, estes definem fundamentos que embasam a representação do conhecimento, domínios específicos e conceituam a resolução de problemas que não dependem do domínio o qual aconteçam, respectivamente. Neste tipo de ontologia, elas também podem ser classificadas quanto ao grau de formalidade que podem ser: altamente informal, é expressa de maneira livre em linguagem natural; estruturada informal, encontrada em linguagem natural mas expressa de maneira restrita; semiformal, definida formalmente e é expressada em uma linguagem artificial; e rigorosamente formal, é expressa por meio de semântica formal, teoremas e provas.

2.2.1 OWL

Desenvolvida pelo W3C a linguagem de Web Semântica conhecida como *Ontology Web Language* (OWL) tem como intuito resolver as limitações encontradas com RDF e RDF *Schema*, desenvolvida baseada nessas duas linguagens e DAML+OIL.

> O DAML+OIL (DARPA Agent Markup Language – Ontology Interchange Language) é uma linguagem baseada no XML, desenhada para possuir muito mais capacidade que este na descrição de objetos e no seu relacionamento; para expressar semântica e criar um alto grau de interoperabilidade entre sites Web (Souza e Alvarenga, 2004:137).

Por ter sido baseada em linguagens que foram construídas em cima do XML, também a possui como base e tem como finalidade atender aos requisitos da Web Semântica disponibilizando algumas características com melhor descrição quando acontece o relacionamento e as definições dos e entre os recursos. Pelo fato de oferecer a criação de um vocabulário adicional para descrição de propriedades e classes possui a expressividade necessária para representar ontologias mais complexas.

De acordo com Moraes (2007) a linguagem OWL consegue suprir as restrições existentes para RDF e RDF *Schema* como: método usado para indicar que os valores de determinada classe são instâncias de uma ou mais classes gera limitação desses valores que podem ser aplicados a uma certa propriedade; a não identificação de classes que tem uma ligação com uma mesma subclasse sejam distintas; não conseguir criar classes a partir de outras usando operações booleanas como intersecção, união e complemento; não oferecer suporte para realizar a definição da quantidade de valores que uma propriedade pode ter; e não conseguir rotular as propriedades como transitivas, únicas ou inversa de outra, fazendo com que não consiga ser aplicada a dedução a partir dos indícios sobre as classes de acordo com suas propriedades.

Para inteirar outras restrições existentes, a linguagem OWL possui três sublinguagens que foram construídas com a derivação da sua antecessora, são essas:

- A OWL-Lite é a versão mais simplificada e tem seu propósito na descrição de restrições e da hierarquia de classes simples, esta é mais simples de

implementar e consequentemente possui melhor desempenho mas tem pouca expressividade da linguagem.

- OWL-DL baseia-se em lógica descritiva o que adiciona a possibilidade de raciocínio automatizada, impondo restrições quanto ao uso dos recursos e melhorando a expressividade da linguagem.
- OWL-Full que para a utilização desta, não é possível realizar deduções em uma ontologia; essa linguagem também não impõe restrição sintática e garante que qualquer documento RDF que seja válido é um documento OWL-Full válido.

2.3 Fusão de Dados

Na fusão de informação, os dados encontrados em um determinado cenário são correlacionados e unidos com auxílio de uma ontologia específica, assim podem ser persistidos de maneira padronizada, fazendo com que o usuário final, seja ele máquina ou não, possa utilizar da informação conforme desejado.

Segundo Botega (apud Pereira, 2016: 19), fusão de dados e informações é a rotina de transformação de dados e informações com a finalidade de produzir estimativas e predições de estados de entidades, tendo como o objetivo a maximização do valor da informação que é adquirida e o estímulo da consciência situacional de analistas em relação a um ambiente desejado.

Pelo fato dos dados serem encontrados de maneira descentralizada e sem uma padronização pré-definida na WEB a fusão de informação nesse cenário pode utilizar de Agentes de Extração de dados juntamente com uma ontologia para realizara extração de informações neste meio, conseguindo unir dados que tenho real valor para o usuário final.

A informação extraída com o uso da união dos Agentes de extração, que nesse trabalho são o Governo, Instituições Educacionais e Empresas, e da ontologia é verificada por um algoritmo que tem como funcionalidade, observar se a informação que foi retirada do cenário WEB está de acordo com o contexto pedido pela ontologia usada e se será útil para quem for usá-la, assim sendo persistida.

Bitencort Junior (2008) fala que a fusão de dados é a capacidade que os sensores computacionais possuem para juntar os dados que foram coletados, consequentemente reduzindo a quantidade de informações e o tamanho destas mensagens que trafegam pela Web.

2.4 JSOUP

Conforme Hedley (2016), Jsoup é uma biblioteca java usada para trabalhar com HTML e fornece uma API que usa do CSS, do *Document Object Model* (DOM) e de métodos *jQuery-like* para trabalhar com HTML para realizar a extração e manipulação de dados. Este implementa a especificação para Html5 (WHATWG HTML5) e é capaz de analisar o HTML para o mesmo DOM como fazem os navegadores mais novos e criar uma árvore construída com uma análise sensata.

Essa biblioteca possibilita algumas funcionalidades como: apurar e analisar o HTML de uma URL, um arquivo ou *string*; manipular elementos, atributos e textos HTML; limpar o conteúdo enviado pelo usuário de encontro a uma lista vazia segura, com o intuito de evitar ataques XSS que é uma vulnerabilidade causada por falha na validação de parâmetros de entrada de um usuário e resposta do servidor na aplicação.

2.5 RDF

Segundo Ferreira e Santos (2013), na década de 1990 foi criado um grupo de trabalho, pelo W3C, intitulado como *Resource Description Framework* (RDF) que buscava discutir uma estrutura de recursos que atingisse as necessidades de diferentes comunidades de descrição que se interessassem, pois percebeu-se que apenas a classificação e a descrição do conteúdo de páginas da Web realizado pelo padrão *Platform for Internet Content Selection* (PICS) era insuficiente, contendo limitações nas especificações.

De acordo com a especificação do W3C (1999), o RDF tem como fundamentação o processamento de metadados que fornece interoperabilidade entre as aplicações que trocam informações e podem ser compreendidas por máquinas, realça facilidades para realizar o processamento automatizado de recursos na Web e realiza o uso do padrão XML para especificação da semântica dos dados; realizando, assim, apenas a representação dos metadados sobre os recursos.

A atualização de 2004 pelo W3C trouxe a possibilidade de descrição dos recursos encontrados na Web, representa os metadados, que são conjunto de atributos e informações sobre os dados referidos no *World Wide Web*, também pode ser usado para representar todas as informações ou objetos que podem ser identificados neste meio, mesmo quando esses dados não podem ser recuperados.

> O RDF é destinado para aquelas situações que as informações precisam ser processadas por aplicativos, e que não são apenas exibidas para pessoas. RDF fornece uma estrutura comum para que essas informações sejam expressas e possam realizar intercâmbios entre aplicações sem perder significado. [. . .] A capacidade de trocar informações entre aplicações diferentes mostra que esta informação pode estar disponível para finalidades diferentes da qual foi gerada inicialmente. (W3C, 2004)

Esta ferramenta é uma estrutura para expressar recursos, estes sendo documentos, pessoas, objetos físicos, conceitos abstratos e outras informações encontradas na Web. Pode ser usado para publicar e interligar dados neste cenário, o que permite que uma pessoa ou processo automatizado possa seguir essas ligações e agregar os dados sobre estes recursos. O RDF é capaz de realizar ligação dos dados que fazem parte de uma organização, e permite consultas cruzadas deste conjunto de dados com utilização do SPARQL; enriquece o *dataset* por meio de interligação com conjunto de dados que possuem recursos vinculados.

A sintaxe deste modelo de dados pode ser realizada por declarações, conhecidas como Triplas compostas por propriedade, recurso e valor que permitem declarações sobre recursos; conforme é mostrado na Figura 1, essas declarações expressam a relação entre dois recursos, um sujeito e um objeto.

Segundo Santarém (apud Coneglian, 2014) o recurso, ou sujeito, é todo elemento que tenha identidade, pode ser serviço, imagem e outros, nem todo ele pode ser recuperável já que pode ser identificado por uma URI e é considerado o mapeamento conceitual para um conjunto de entidades ou uma única, e a propriedade, por sua vez, possui um respectivo valor (objeto) e caracteriza um recurso, um único recurso pode ter mais de uma propriedade.

Dias e Santos (2003) falam que a função exercida pelo Esquema RDF ou RDFS é “permitir a criação de classes de tipos de recursos e propriedades, descrições dessas classes, combinações possíveis de classes, propriedades e valores e restrições entre relacionamentos, definindo assim esquemas que podem ser utilizados em conjunto com

vocabulários descritivos".

Para ser utilizada a API conhecida como Jena necessita que o desenvolvedor tenha familiaridade tanto com a linguagem de programação Java como com XML. Possui métodos para ler e escrever RDF e XML, e é capaz de armazenar um RDF em um arquivo e carregá-lo em outro.

2.6 SPARQL

De acordo com Elias e Holanda (2016) o *Simple Protocol and RDF Query Language* é uma linguagem construída para consulta da web semântica, capaz de permitir a recuperação de valores de dados estruturados e semiestruturados, a exploração dos dados quando realizada consultas com relações desconhecidas e uniões complexas de diferentes *datasets* em uma consulta única e simples.

Conforme a recomendação disponibilizada em Janeiro de 2008 pelo W3C essa linguagem de consultas contém capacidades necessárias para requisitar o que foi solicitado, pode ser usada para expressar buscas através de dados com diferentes fontes, estando eles armazenados nativamente como RDF ou vistos como RDF por meio de *middleware* e seus resultados podem ser tanto um conjunto de resultados quanto de gráficos RDF. Realiza consultas de padrões em RDF, as buscas são enviadas por HTTP e os seus resultados podem ser disponibilizados tanto em formato XML como em formato JSON.

Nas versões atualizadas da consulta e do protocolo SPARQL foi especificado pela W3C em março de 2013 a possibilidade: de inserção, remoção a modificação dos dados RDF por meio do SPARQL1.1 *Update*; combinação de inferência pelo *Entailments*; consulta de uma única vez de muitos *endpoints*; resultados com formato de CSV/TSV e outros.

Elias e Holanda (2016) dizem que a estrutura desta consulta é composta por declarações de prefixos que tem o intuito de abreviar URIs, definição do conjunto de dados informando os grafos RDF que são consultados, identificação da informação que deve retornar a partir da consulta pela cláusula de resultado, o padrão de consulta que está sendo usado que indica o que deve ser consultado dentro do conjunto de dados e os modificadores de consulta, limites, ordenação, e outros que tem poder de modificar o resultado final da busca.

3. Arquitetura Semântica de Extração Informacional

Já que não existe um padrão específico para os dados que são colocados no ambiente Web, são encontrados dados de diferentes lugares e com diversos formatos e estruturas, muitos destes podem ser semiestruturados ou até mesmo não-estruturados, não possuindo um formato padrão previamente definido.

Em processos de Recuperação da Informação, uma alta quantidade de informações e dados podem não tem relevância real, sendo o próprio usuário o responsável por verificar as informações de maior importância.

Nesta seção apresenta-se a arquitetura de referência do projeto para extração de dados na Web, esta é composta pelo conjunto dos elementos que participam de toda a estrutura. Essa arquitetura pode ser subdividida em duas, como mostrado na Figura 1, são essas subdivisões: os Atores de Inovação e o Espaço Informacional que será utilizado.

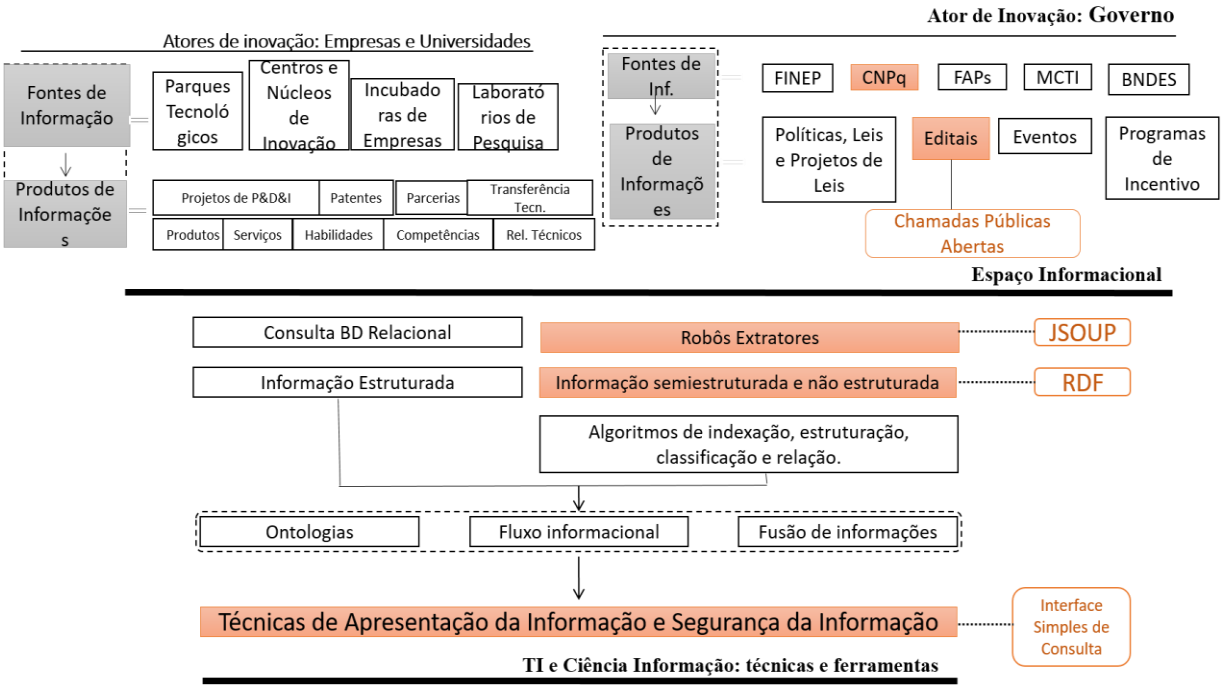
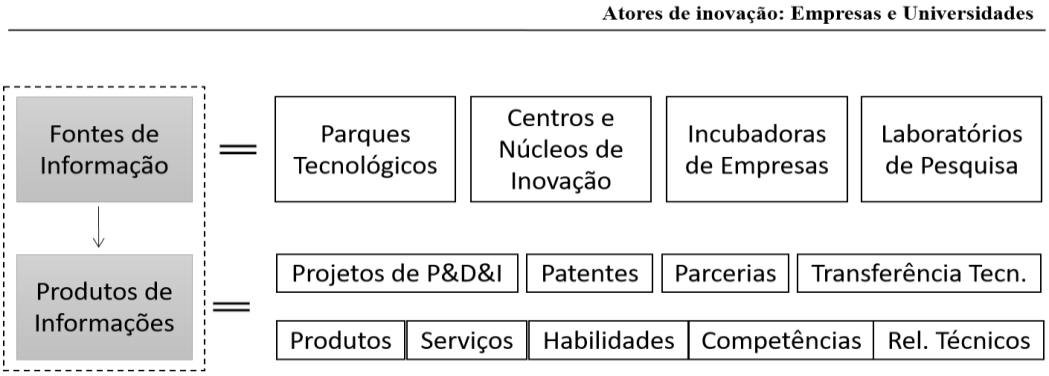


Figura 1 - Arquitetura Semântica de Extração Informacional
Fonte: Autoria Própria

A Figura 1 mostra as Fontes e os Produtos de Informações dos Atores de Inovação Tecnológica, e no Espaço Informacional é encontrado o modelo com os elementos necessários para a construção de um Robô Extrator. A arquitetura demonstrada na Figura 1 mostra o que é utilizado para realizar a extração das informações que são usadas na construção da estrutura RDF, montada com auxílio do *Framework* Jena.

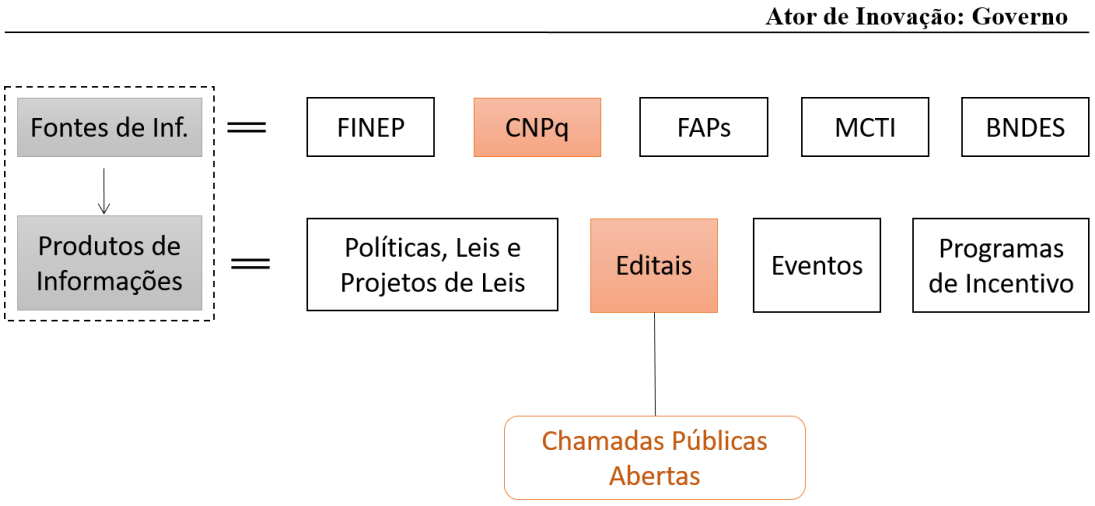


Os Produtos de Informações são os objetos produzidos pelas Fontes de Informação como mostram as Figuras 2 e 3, estas mostram os Atores de Inovação: Empresas, Universidades e Governo.

Figura 2 - Estrutura dos Atores de Inovação: Empresas e Universidades
Fonte: Autoria Própria

Pode ser observada na Figura 3 que com relação ao Governo, as informações

que foram extraídas são obtidas pelo Edital de Chamadas Públicas Abertas disponibilizado pelo CNPq.



Neste estudo apenas alguns destes elementos são utilizados, já que as informações extraídas podem ser semiestruturadas ou não-estruturadas, dependendo do local da extração.

A Figura 4 mostra que o Robô Extrator é desenvolvido com auxílio da biblioteca Java Jsoup, capaz de realizar a leitura das informações semiestruturadas encontradas e retiradas da página do Cnpq e estas são modeladas pelas triplas de RDF, com auxílio do *Framework* Jena e são apresentadas para o usuário final por meio de uma interface simples.

O Espaço Informacional é composto por todos os elementos existentes para a realização da Extração de Informações. Para este projeto, é realizada a extração das informações desejadas com auxílio do JSOUP; a estruturação dessas informações no formato RDF; a consulta que é realizada pela linguagem SPARQL e a visualização das informações encontradas que são solicitadas pelo usuário.

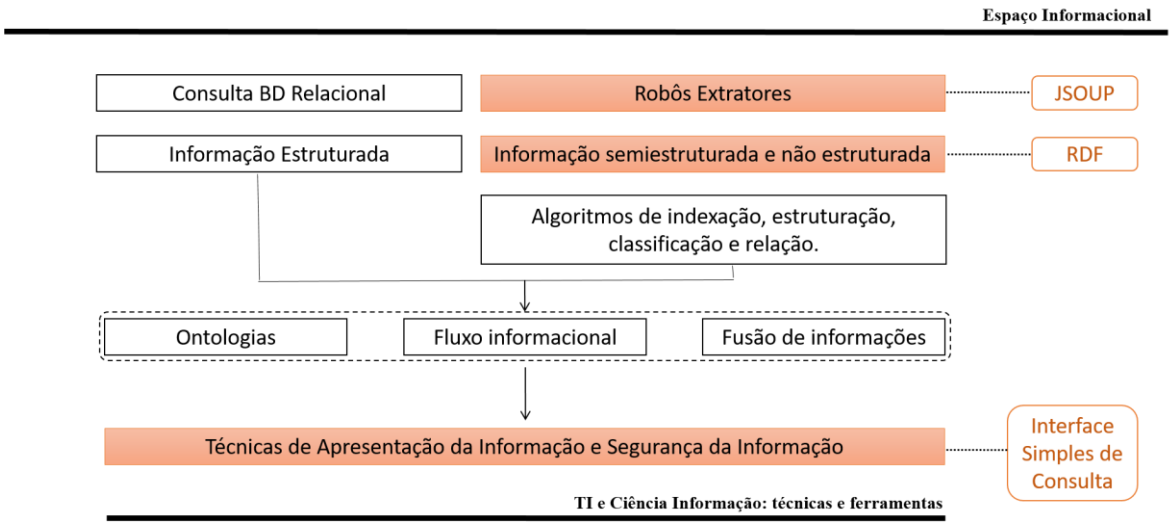


Figura 4 - Estrutura do Espaço Informacional
Fonte: Autoria Própria

É observado que para a automação dos Robôs Extratores, desde o momento da leitura das páginas Web até o momento da apresentação dessas informações para o usuário final, que os dados recuperados passam por uma série de processos internos até que sejam apresentados ao usuário.

4. Extração e Modelagem dos Dados

A extração e estruturação das informações desejadas é realizada no âmbito do Espaço Informacional. Esta fase do projeto é a responsável pela formatação das informações que são extraídas e utilizadas posteriormente.

Os dados capturados são retirados da página de Chamadas Públicas Abertas do CNPQ e estão presentes de forma semiestruturada, já que ela possui um formato de apresentação para o usuário que é pré-definido.

O diagrama apresentado na Figura 5, a seguir, mostra o caminho que é percorrido por uma determinada informação quando realizada uma consulta pelo usuário do sistema.

A Figura 5 mostra que os dados obtidos são encontrados na página de Chamadas Públicas Abertas do CNPQ. Esses valores encontrados na Web passam pelo Processo de Extração e Modelagem, seguido pelo Processo de Consultas, até o determinado momento da visualização da informação que é solicitada pelo usuário.

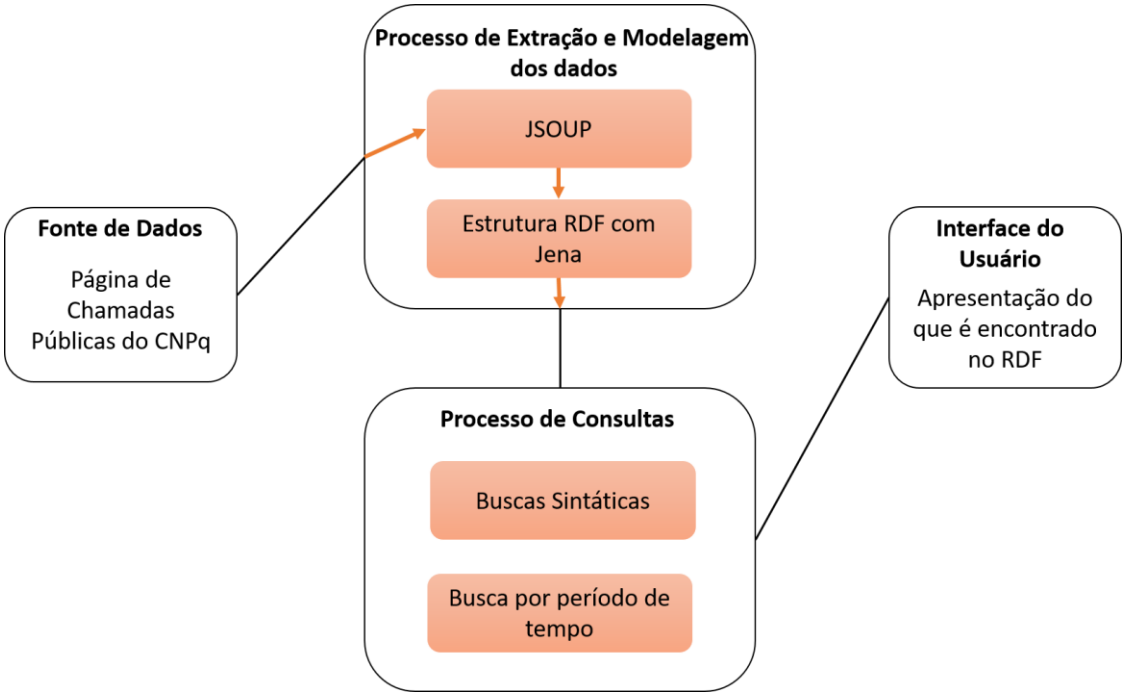


Figura 5 - Diagrama do Processo de Extração e Modelagem
Fonte: Autoria Própria

A Figura 6 exemplifica, de maneira genérica, como foi realizada a implementação da primeira estrutura RDF construída. Esta possuía um atributo genérico para chamada que era ligado por triplas com os elementos título, descrição, link e data que são guardados. Todos os valores que deveriam ser atribuídos a cada um dos atributos eram separados pela sua classificação e a ligação realizada entre o atributo e o valor era realizado pelo respectivo link que foi extraído.

Pelo fato dessa estrutura mostrada na Figura 6 ser inviável para a realização das consultas que são realizadas, foi definido e implementado um novo modelo de RDF, conforme exemplificado na Figura 7.

De cada item existente na lista são retiradas as informações de link, data completa, título e descrição de cada chamada aberta disponibilizada. Para cada chamada existente nesta lista é gerada uma estrutura de triplas RDF que contém as informações que foram recuperadas e os elementos de Data Inicial e Data Final da chamada, como demonstrado na Figura 7.

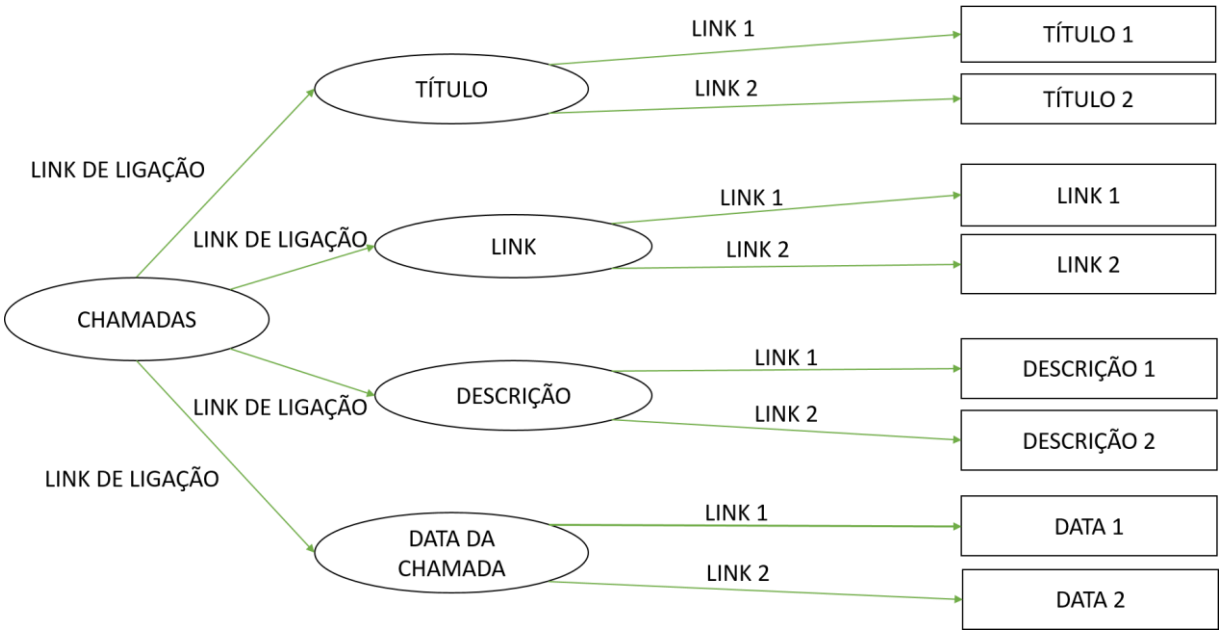
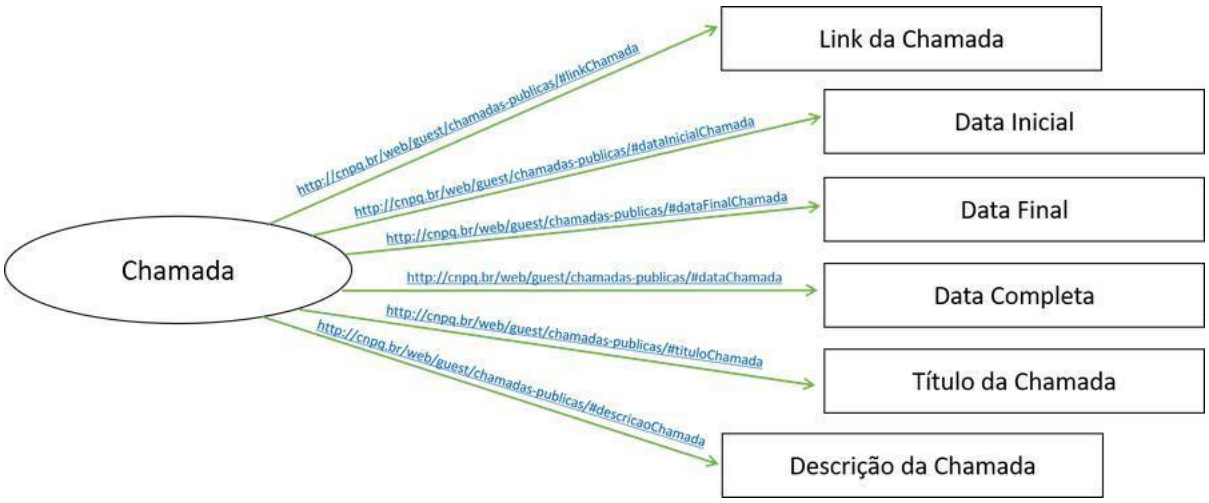
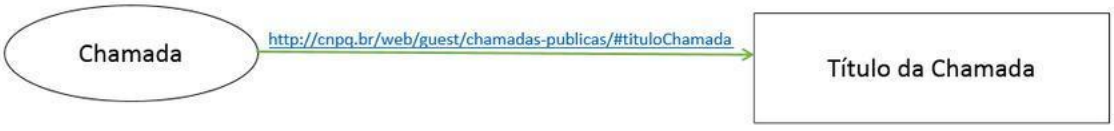


Figura 6 - Exemplificação Estrutura RDF Inicial
Fonte: Autoria Própria



com o auxílio da biblioteca Jena, usada na linguagem Java. O recurso Chamada possui os literais (valores) de link, data inicial, data final, data completa, título e descrição que são conectados por links estáticos e específicos para cada valor, com o intuito de realizar a consulta das informações posteriormente.

Na Figura 8 é apresentado um exemplo da estrutura de uma das triplas RDF, existente para cada recurso, que são construídas para estruturação das informações desejadas.



Para realizar a obtenção das informações desejadas é realizada uma conexão com a página especificada, esta ligação é realizada através de um método existente e disponibilizado pela biblioteca Jsoup.

Além da conexão com a página Web, o Jsoup é capaz de realizar uma varredura ampla e capturar qualquer informação que nela esteja disponibilizada. Por possuir métodos específicos para tais tarefas, foi possível especificar os elementos da lista principal que deveriam ser capturados, usando principalmente o método getElementsByClass(), já que esses elementos são encontrados em classes específicas existentes no Html.

Após a extração das informações da página do Cnpq, essas passam pelo processo de inserção no RDF que já teve suas propriedades e recursos criados anteriormente. Esse procedimento acontece com a adição dos valores dessas propriedades com auxílio do addProperty(), método do *Framework* aberto Jena.

Desse modo, foi construída a estrutura RDF a qual as informações capturadas são inseridas, possibilitando a realização das consultas realizadas de acordo com a necessidade do usuário.

5. Agente de Consulta

Assim como a extração das informações e a esquematização destas, as consultas feitas em cima da estrutura criada são realizadas na parte de Espaço Informacional deste projeto.

Durante o processo de decisão foi definido que as consultas que seriam realizadas pelo usuário são capazes de extrair informações, contidas no RDF, baseadas em palavras e num determinado período de tempo.

Para fazer as consultas desejadas é usada a linguagem para consultas de Web Semântica conhecida como SPARQL que é disponibilizada pelo framework Jena. As *querys* usadas por essa linguagem fazem as buscas especificadas em tripas RDF e para que sejam retornadas as informações encontradas nessas estruturas é necessário que estejam modeladas contendo as especificações separadas por módulos.

De acordo com o a Figura 9, é necessário que o item de retorno desejado seja referenciado pelo link de ligação, ou seja, a propriedade usada para ligação entre o recurso e o valor deve ser usado para que o SPARQL consiga realizar e retornar os valores encontrados durante a busca na estrutura.

```
String queryStr
= "SELECT ?title ?descricao ?link ?data WHERE "
 + "{?x <http://cnpq.br/web/guest/chamadas-publicas/#tituloChamada> ?title FILTER (regex(?title, '"+testeTitulo1+"', 'i')   "
 + " || regex(?title, '"+testeTitulo2+"', 'i') "
 + " ). "
 + "?x <http://cnpq.br/web/guest/chamadas-publicas/#descricaoChamada> ?descricao . "
 + "?x <http://cnpq.br/web/guest/chamadas-publicas/#linkChamada> ?link . "
 + "?x <http://cnpq.br/web/guest/chamadas-publicas/#dataChamada> ?data }";
```

No *select* construído para realização da leitura por palavras é necessário o uso de um *filter* com a função *regex*. A expressão construída varre todo o RDF à procura dos elementos que são solicitados na busca feita pelo usuário pela interface. Para cada palavra informada deve ser construído e adicionado uma nova função *regex* no *filter* feito para consulta.

Já para as pesquisas realizadas pelas datas é necessário que os elementos usados na leitura das informações estejam em um formato específico, como mostrado na Figura 10.

Para essa consulta, a expressão implementada contém apenas a função *filter*, nela foi feita uma comparação entre as datas com o intuito de recuperar todas as chamadas públicas abertas existentes no intervalo de tempo especificado e que foram inseridas na estrutura RDF.

A implementação das consultas que são realizadas pelo SPARQL mostra que, para recuperar as informações corretas e desejadas, os agentes de busca utilizados devem estar previamente formatados, de acordo com o que é aceitado pela linguagem e dentro das funções especificadas para cada tipo de consulta que pode ser efetuada.

```
String queryDataUma
                = "SELECT ?dataFim ?dataInicio ?title ?descricao ?link ?data WHERE "
                        + "{?x <http://cnpq.br/web/guest/chamadas-publicas/#dataFinalChamada> ?dataFim ; "
                        + " <http://cnpq.br/web/guest/chamadas-publicas/#dataInicialChamada> ?dataInicio . "
                        + " FILTER ( ?dataFim >= '2016-12-17' || ?dataInicio <= '2016-09-18' ) . "
                        + "?x <http://cnpq.br/web/guest/chamadas-publicas/#tituloChamada> ?title . "
                        + "?x <http://cnpq.br/web/guest/chamadas-publicas/#descricaoChamada> ?descricao . "
                        + "?x <http://cnpq.br/web/guest/chamadas-publicas/#linkChamada> ?link . "
                        + "?x <http://cnpq.br/web/guest/chamadas-publicas/#dataChamada> ?data }";
```

Figura 10 - Exemplo da Consulta por Data
Fonte: Autoria Própria

6. Resultados

Para verificação das informações que são encontradas de acordo com as consultas, dois tipos de pesquisa foram realizadas: uma busca baseada em palavras desejadas e outra para um determinado período de tempo.

Atualmente, estão disponibilizadas poucas chamadas na página web do CNPq selecionada. Com isso, a estrutura RDF foi gerada com 2 itens, estes compostos com todas as informações desejadas.

O exemplo demonstrado na Tabela 1 representa a consulta por palavras. Nesse caso, a busca realizada foi pela palavra ‘apoio’.

Tabela 1 - Verificação das Chamadas Recuperadas quando a leitura por palavra

Título	Descrição	Período da Chamada	Possui a palavra solicitada?
Chamada CNPq/MCTIC Nº 016/2016 - SEGURANÇA ALIMENTAR E NUTRICIONAL NO ÂMBITO DAUNASUL	Objeto Apoiar projetos de pesquisa científica e tecnológica que visem contribuir significativamente para o desenvolvimento científico (...)	20/09/2016 a 16/12/2016	Não
APOIO ÀPESQUISA E ÀINOVAÇÃO EM CIÊNCIASHUMANAS,SOCIAIS ESOCIAIS APLICADAS	A presente chamada tem por objetivo apoiar atividades de pesquisa de excelência, inovadoras e criativas, nos temas elencados nas Linhas de Pesquisa, com foco (...)	12/09/2016 a 23/01/2017	Sim

Na tabela 1, são apresentadas as informações de título, descrição e data de todas as chamadas que foram extraídas da página do CNPq e são encontradas no RDF, também pode ser encontrado se a chamada pública em questão possui a palavra usada na consulta.

As informações apresentadas na Tabela 2 são referentes a uma consulta que é realizada com um conjunto de palavras. Neste exemplo, a consulta foi realizada pelas palavras ‘apoio’ e ‘segurança’.

Tabela 2 - Verificação das Chamadas Recuperadas quando a leitura por um conjunto de palavras

Título	Descrição	Período da Chamada	Possui alguma das palavras solicitadas?	Palavra Encontrada
Chamada CNPq/MCTIC Nº 016/2016 - SEGURANÇA ALIMENTAR E NUTRICIONAL NOÂMBITO DAUNASUL	Objeto Apoiar projetos de pesquisa científica e tecnológica que visem contribuir significativamente para o desenvolvimento científico (...)	20/09/2016 a 16/12/2016	Sim	Segurança
APOIO À PESQUISA E À INOVAÇÃO EMCIÊNCIAS HUMANAS, SOCIAIS E SOCIAIS APLICADAS	A presente chamada tem por objetivo apoiar atividades de pesquisa de excelência, inovadoras e criativas, nos temas elencados nas Linhas de Pesquisa, com foco (...)	12/09/2016 a 23/01/2017	Sim	Apoio

A Tabela 2 apresenta as informações de título, descrição e data de todas as chamadas que foram extraídas da página do CNPq e são encontradas no RDF, assim como se a chamada pública em questão possui alguma das palavras que foi usada na consulta e qual a palavra encontrada em cada título que foi retornado.

Para que as chamadas sejam mostradas na interface, quando é realizada uma consulta por um intervalo de tempo, é preciso que elas estejam no intervalo de tempo informado pelo usuário.

Para a demonstração da busca apresentada na tabela abaixo é usado o intervalo de tempo que possui data inicial em 23/09/2016 e 18/12/2016.

Tabela 3 - Verificação das Chamadas Recuperadas por consulta por intervalo de tempo

Título	Descrição	Período da Chamada	Presente no Intervalo?
Chamada CNPq/MCTIC Nº 016/2016 - SEGURANÇA ALIMENTAR E NUTRICIONAL NO ÂMBITO DAUNASUL	Objeto Apoiar projetos de pesquisa científica e tecnológica que visem contribuir significativamente para o desenvolvimento científico (...)	20/09/2016 a 16/12/2016	Sim
APOIO À	A presente chamada tem	12/09/2016 a	Sim

PESQUISA E À INOVAÇÃO EM CIÊNCIAS HUMANAS,SOCIAIS ESOCIAIS APLICADAS	por objetivo apoiar atividades de pesquisa de excelência, inovadoras e criativas, nos temas elencados nas Linhas de Pesquisa, com foco (...)	23/01/2017	

A tabela 2 mostra as mesmas informações da tabela 1 quanto ao que é apresentado sobre as chamadas, porém, a coluna relacionada a consulta exemplifica a leitura realizada no intervalo de tempo pré-definido.

No decorrer da implementação do projeto, foram encontradas diversas dificuldades, principalmente relacionadas as consultas que deveriam ser realizadas. Com essas dificuldades, pôde-se observar que a consulta realizada com a linguagem SPARQL apenas é capaz de retornar os valores corretos se a estrutura montada para guardar as informações desejadas seja encontrada no padrão mostrado na Figura 7.

Assim, é entendido que para as informações das chamadas serem retornadas de acordo com a busca realizada por um período de tempo é preciso que elas estejam, em algum momento, dentro do intervalo que foi solicitado. Caso dentro do período pedido não tenha informações de chamadas abertas, nenhum item será recuperado.

O esboço implementado neste estudo não possui a integração entre as informações que são extraídas da página por meio da Fusão de Informações, assim como também não são realizadas buscas observando a relação semântica das palavras consultadas. Também não foi realizada a modelagem de apresentação das informações, sendo, estes apresentados para o usuário em uma interface simples, sem a informação de geolocalização dos dados.

Este trabalho é parte de um projeto geral que é composto em diversas etapas. O presente esquema tem como função contribuir com a validação de tecnologias, como o Jena e Jsoup, que podem ser usadas no processo de extração e armazenamento dos dados que são usados por mecanismos no momento da recuperação pela linguagem SPARQL.

Conclusões

O modelo proposto de extração baseado em tecnologias semânticas serve de base para construção de um ambiente de Web Semântica com a utilização de ontologias de domínio da área da inovação. Os dados recuperados não possuem uma verificação baseada no significado e no contexto de termos buscados, sendo o usuário o responsável por decidir a relevância das informações apresentadas na interface.

Esse projeto teve como objetivo o desenvolvimento de um robô de extração de dados semiestruturados capaz de extrair, estruturar e filtrar informações encontradas em uma página Web, utilizando bibliotecas e softwares como Jena, Jsoup e SPARQL, com o intuito de traçar um modelo que melhore a extração e esquematização destes dados de forma a maximizar a relevância nos ambientes de recuperação da informação.

Como forma de validar o modelo, foi usada a estrutura em RDF, preenchida com as informações das chamadas públicas abertas. O autômato construído se depara com o problema de captura das informações ao encontrá-las semiestruturadas e em alguns casos não-estruturadas.

No momento em que a estrutura demonstrada na Figura 8 estava sendo utilizada, não era obtido nenhum tipo de retorno, fosse ele com resulta ou ocorrência de

erro, nas tentativas de consultas que eram realizadas. Mostrando que existe uma estrutura padronizada para que o SPARQL seja capaz de realizar as consultas sob as informações que são salvadas no RDF.

Para a realização dos testes, o robô de buscas foi implementado com a capacidade de extrair informações da página do CNPq e a consulta que pode ser realizada foi construída utilizando funções disponibilizadas pelo SPARQL. Para que a leitura do RDF aconteça da maneira correta, é necessário que tanto a estrutura gerada como os dados enviados pelo usuário, estejam construídos e formatados dentro dos padrões correspondentes para cada tipo de consulta que é construída.

Depois da realização dos testes, foi observado que o uso da consulta para a estrutura gerada é uma maneira eficaz para se obter informações pesquisadas, atendendo o que foi solicitado pelo usuário.

Grande parte das páginas são criadas para serem lidas apenas pelo usuário, sem uma estrutura e formato que agentes computacionais consigam realizar a extração dos dados ali contidos dentro de um contexto. Assim, pode ser concluído que o uso de uma ontologia é de grande importância para que o autômato realize as tomadas de decisões de acordo com valor semântico dos dados que são recuperados, caso aquela informação esteja dentro do contexto desejado pelo usuário.

O agente de extração construído realiza a leitura e captura dos elementos das chamadas públicas abertas disponibilizadas no site do CNPq que devem ser guardados, enquanto, a implementação realizada com o Jena é capaz de estruturar esses dados no RDF e disponibiliza a linguagem SPARQL para fazer a busca desejada pela estrutura, e assim apresentar os resultados encontrados na interface.

Desta forma, os resultados obtidos com a utilização do protótipo desenvolvido conseguem apresentar aos usuários os dados, filtrados pelas consultas por palavras e datas, que são obtidos de uma página semiestruturada.

Portanto, a recuperação da informação ocorre de maneira sintática, e a partir do que foi extraído, pode ocorrer uma análise das informações baseada na semântica no momento que for inserido o uso de uma ontologia nesse processo. Este método se mostrou muito eficiente, pois consegue realizar a extração e estruturação dos dados da página do CNPq e consegue fazer uma busca observando a sintaxe dos dados, assim apresenta para o usuário aquelas chamadas que possuem a palavra desejada no título e/ou que estão no período de tempo solicitado.

Referências

APACHE. The core RDF API. Acessado em Mai 03, 2016. Disponível em: <https://jena.apache.org/documentation/rdf/index.html>.

APACHE. Uma introdução a RDF e à API RDF de Jena. Acessado em Mai 02, 2016. Disponível em: <https://jena.apache.org/tutorials/rdf_api_pt.html>.

BRÄSCHER, M. WEB SEMÂNTICA. 2007. Acessado em Abr 31, 2016. Disponível em: <http://www.stf.jus.br/arquivo/sijed/16.pdf>.

CONEGLIAN, C. S. Agente Semântico de Extração Informacional no Contexto de Big Data. Marília, 2014. Acessado em Nov 13, 2016. Disponível em: <http://aberto.univem.edu.br/bitstream/handle/11077/997/Caio%20Saraiva%20Coneglian.pdf?sequence=1>.

DIAS, T. D.; SANTOS, N. Web Semântica: Conceitos Básicos e Tecnologias Associadas. Cadernos do IME: Série Informática, Rio de Janeiro, v. 14, p. 79 – 92, Junho 2003. Acessado em Mai 02, 2016. Disponível em: <http://www.e-

publicacoes.uerj.br/index.php/cadinf/article/viewFile/6619/4734>.
ELIAS, E.; HOLANDA, O. SPARQL: Linguagem de Consulta em Ontologias. 2016. Acessado em Mai 13, 2016. Disponível em: <http://www.egov.ufsc.br/portal/sites/default/files/sparqlrevisado.pdf>.
FERREIRA, J. A.; SANTOS, P. L. V. A. da C. O MODELO DE DADOS RESOURCE DESCRIPTION FRAMEWORK (RDF) E O SEU PAPEL NA DESCRIÇÃO DE ECURSOS. Informação & Sociedade: Estudos, João Pessoa, v. 23, n. 2, p. 13 – 23, maio/agosto 2013. ISSN 1809-4783. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/15436/9681>.
FREITAS, F. L. G. Ontologias e a Web Semântica. Acessado em Abr 31, 2016. Disponível em: <http://www.inf.ufsc.br/~fernando.gauthier/EGC6006/material/Aula%203/Ontologia_Web_semantica%20Freitas.pdf>.
HEDLEY, J. Jsoup: Java HTML Parser. 2009-2016. ed. [S.l.], 2016. Acessado em Abr 27,2016. Disponível em: <https://jsoup.org/>.
BITENCORTJUNIOR, B. R. FUSÃO DE DADOS PARALELA EM REDES DE SENSORES SEM FIO DENSAS UTILIZANDO ALGORITMO GENÉTICO. Florianópolis 2008. Acessado em Mai 01, 2016. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/90842/254976.pdf?sequence=1>.
MORAES, M. O. REENGENHARIA DO *ONTOCOVER*: UMA BIBLIOTECA JAVA PARA MANIPULAR ONTOLOGIAS EM APLICAÇÕES DA WEB SEMÂNTICA. Florianópolis 2007. Acessado em Mai 03, 2016. Disponível em: <https://projetos.inf.ufsc.br/arquivos_projetos/projeto_574/tccMarcelo.pdf>.
NAVARRO, M. B. M. de A. et al. Inovação Tecnológica e as questões reflexivas do campo da biossegurança. *Estudos Avançados*, São Paulo, v. 28, n. 80, Jan/Apr 2014. ISSN 0103-4014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142014000100019>
OLIVEIRA, A. B. F. de; WERNECK, V. M. B. Ontologias. Cadernos do IME: Série Informática, Rio de Janeiro, v. 15, p. 7 - 13, dezembro 2003. ISSN: 1413-9014. Disponível em: <http://www.e-publicacoes.uerj.br/index.php/cadinf/article/viewFile/6384/4547>.
PEREIRA, F. D. Automação do fluxo Informacional entre atores de inovação no Brasil para processos de tomada de decisão. Acessado em Fev 08, 2016. 2016.
PRYBECZ, B. H.; GONÇALVES JÚNIOR, J. I.; MENDES, T. R. Web Semântica. Curitiba, 2013. Acessado em Mai 01, 2016. Disponível em: <http://www.inf.ufpr.br/bmuller/TG/TG-BTJ.pdf>.
São Paulo. *DECRETO Nº 56.424, DE 23 DE NOVEMBRO DE 2010*. Acessado em Fev 17, 2016. Disponível em: <http://dobuscadireta.imprensaoficial.com.br/default.aspx?DataPublicacao=20101124&Caderno=DOE-I&NumeroPagina=1>.
São Paulo. *DECRETO Nº 60.286, DE 25 DE MARÇO DE 2014*. Acessado em Fev 17, 2016. Disponível em: <http://www.al.sp.gov.br/repositorio/legislacao/decreto/2014/decreto-60286-25.03.2014.html>.
SECRETARIA DE DESENVOLVIMENTO ECONÔMICO, CIÊNCIA, TECNOLOGIA E INOVAÇÃO. *CENTRO DE INOVAÇÃO*. Acessado em Fev 16, 2016. Disponível em: <http://www.desenvolvimento.sp.gov.br/centros-de-inovacao>.
SECRETARIA DE DESENVOLVIMENTO ECONÔMICO, CIÊNCIA, TECNOLOGIA E INOVAÇÃO. *REDE PAULISTA DE INCUBADORAS*. Acessado em Fev 16, 2016. Disponível em: <http://www.desenvolvimento.sp.gov.br/centros-de-inovacao>.

SILVA, G. C. RDF e RDFS na Infra-estrutura de Suporte à Web Semântica. Acessado em Mai 04, 2016. Disponível em: <http://www2.ic.uff.br/~gsilva/slreic.pdf>.
SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. Ciência da Informação, Brasília 33.1. 132-141. 2004. Acessado em Mai 02, 2016.
UFBA. Ambiente Virtual de Aprendizagem. Universidade Federal da Bahia, 2007. Acessado em Fev 22, 2016. Disponível em: <http://www.moodle.ufba.br/mod/book/view.php?id=10902&chapterid=9850>.
W3C. RIF RDF and OWL Compatibility (Second Edition). 2003. Acessado em Mai 03, 2016. Disponível em: <https://www.w3.org/TR/2013/REC-rif-rdf-owl-20130205/>.
W3C. Resource Description Framework (RDF) Model and Syntax Specification. 1999. Acessado em Mai 02, 2016. Disponível em: <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
W3C. RDF Primer. 2004. Acessado em Mai 03, 2016. Disponível em: <https://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
W3C. RDF 1.1 Primer. 2004. Acessado em Mai 03, 2016. Disponível em: <https://www.w3.org/TR/rdf11-primer/>
W3C. RDF and SPARQL: Using Semantic Web Technology to Integrate the World's Data. 2007. Acessado em Mai 05, 2016. Disponível em: <https://www.w3.org/2007/03/VLDB/>.
W3C. SPARQL 1.1 Protocol. 2013. Acessado em Mai 06, 2016. Disponível em: <https://www.w3.org/TR/sparql11-protocol/>.
W3C. SPARQL Query Language for RDF. 2008. Acessado em Mai 06, 2016. Disponível em: <https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>.
W3C. SPARQL Protocol for RDF. 2008. Acessado em Mai 5, 2016. Disponível em: <https://www.w3.org/TR/rdf-sparql-protocol/>.