CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA

FUNDAÇÃO DE ENSINO “EURÍPIDES SOARES DA ROCHA”

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PLATAFORMA DE RECUPERAÇÃO SEMÂNTICA E COGNITIVA DE INFORMAÇÕES DE APOIO À INOVAÇÃO

THIAGO APARECIDO GONÇALVES DA COSTA

ORIENTADOR: PROF. DR. FÁBIO DACÊNCIO PEREIRA

Marília - SP
Dezembro/2017

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA

FUNDAÇÃO DE ENSINO “EURÍPIDES SOARES DA ROCHA”

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PLATAFORMA DE RECUPERAÇÃO SEMÂNTICA E COGNITIVA DE INFORMAÇÕES DE APOIO À INOVAÇÃO

THIAGO APARECIDO GONÇALVES DA COSTA

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Eurípides de Marília como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.
Orientador: Prof. Dr. Fábio Dacêncio Pereira

Marília - SP

Dezembro/2017

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA - UNIVEM
MANTIDO PELA FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Thiago Aparecido Gonçalves da Costa

PLATAFORMA DE RECUPERAÇÃO SEMÂNTICA E COGNITIVA DE INFORMAÇÕES DE APOIO À INOVAÇÃO

Banca examinadora da monografia apresentada ao Curso de Bacharelado em Ciência da Computação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Ciência da Computação.

Nota: 10 (Dez)

Orientador:Fábio Dacêncio Pereira

1º.Examinador:Elvis Fusco

2º.Examinador:Caio Saraiva Coneglian

Marília, 01 de dezembro de 2017.

Dedico este trabalho a Deus que me dá forças e capacidade de sonhar e a minha família que me apoia incondicionalmente.

Agradecimento

Meus sinceros votos de agradecimentos,

Primeiramente, agradeço a Deus por ter me dado saúde, oportunidade de estudar e por ser o guia da minha vida.

Agradeço aos meus pais João Cícero G. da Costa e Sueli P. G. da Costa pelo amor, incentivo e apoio incondicional.

Agradeço a minha irmã Karina, meu tio João Pedrozo e minha avó Francisca pelo apoio e incentivo ao meu interesse de continuar os estudos.

Ao meu orientador Prof. Dr. Fábio Dacêncio Pereira por ter sido um amigo nestes quatro anos de graduação, onde me incentivou e oportunizou meu desenvolvimento acadêmico e técnico.

Aos professores Dr. Allan Cesar M. de Oliveira, Dr. Elvis Fusco, Dr. Fábio D. Pereira, Dr. Leonardo C. Botega, Ms. Rodolfo B. Chiaramonte por sempre terem me encorajado a ministrar minicursos, participar e organizar eventos, concorrer no processo seletivo para o programa de bolsas Ibero-Americanas do Santander e despertar o meu interesse pelo Mestrado em Ciência da Computação.

Aos professores Maurício Duarte e Paulo R. M. Cardoso pela didática incrível em sala de aula.

Aos meus amigos e colegas do COMPSI: Cassio Viana, Claudio Costa, Danielle Peraccini, Fernanda Mayumi, Frederico Soares, João Ricardo, Leonardo Ademir, Lívia, Luan William, Marina Beretta, Matheus Ferraroni, Valdir Junior, entre outros.

Ao PROUNI por ter financiado uma parte da minha graduação.

Ao CNPQ e a FAPESP (2016/13025-0) por terem confiado em minha pessoa e oferecido bolsas de iniciação científica que além de financiar os meus projetos aumentaram o meu interesse pela pesquisa e área acadêmica.

Ao Santander Universidades por ter me proporcionado um intercâmbio na Europa que resultou numa experiência acadêmica e cultural fantástica.

Ao UNIVEM pelo suporte e fomento acadêmico no decorrer destes quatro anos de Bacharelado em Ciência da Computação.

RESUMO

Atualmente no domínio da Web com a produção exacerbada de informações, as áreas de Obtenção, Tratamento e Recuperação da Informação estão sendo postas em cheque pelo volume, variedade e velocidade dos dados estruturados e não estruturados de natureza complexa que devem ser encontrados e julgados quanto ao seu valor e sua veracidade, além disso, este fenômeno citado denomina-se Big Data. Desse modo, o processo de inovação tornou-se o foco de muitas empresas para melhorar a sua competitividade, seu posicionamento em novos mercados e para agregar conhecimento mais preciso ao seu modelo de negócio. Portanto, este trabalho tem o objetivo de elaborar uma Plataforma de Apoio à Inovação que tenha a capacidade de Extrair, Classificar e Recuperar informações semânticas e cognitvas, sendo que o espaço informacional será delimitado a atores de inovação (governo, empresas e universidades) do Sistema Paulista de Ambientes de Inovação (SPAI). Ademais, foram obtidos resultados que vão desde a criação da Plataforma de Apoio à Inovação (PLATSINN) até a validação do robô de busca implementado.

Palavras-chave: inovação, extração, web semântica, computação cognitiva, big data

ABSTRACT

Currently in the field of the Web with the exacerbated production of information, the areas of obtaining, treatment and recovery of information are being put into check by volume, variety and speed of structured and unstructured data of a complex nature that should be Found and judged on their value and their veracity, moreover, this phenomenon cited is called Big date. Thus, the innovation process has become the focus of many companies to improve their competitiveness, their positioning in new markets and to aggregate more precise knowledge to their business model. Therefore, this work aims to draw up a platform for innovation support that has the ability to extract, classify and recover semantic information and Cognitvas, and the informational space will be delimited to actors of innovation (government, Companies and universities) of the Paulista System of Innovation Environments (SPAI). Furthermore, results ranging from the creation of the Innovation Support Platform (PLATSINN) to the validation of the search robot implemented.

Keywords: innovation, extraction, semantic web, cognitive computing, big data

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API – *Application Programming Interface*

CIEM – *Centro Incubador de Empresas de Marília*

CNPQ – *Conselho Nacional de Desenvolvimento Científico e Tecnológico*

FAPESP – *Fundação de Amparo à Pesquisa do Estado de São Paulo*

HTML - *HyperText Markup Language*

IBM – *Internation Business Machines*

NoSQL – *Not Only SQL*

OWL – *Web Ontology Language*

PLATSINN – *Platform to Support Innovation*

RDF – *Resource Description Framework*

SQL – *Structured Query Language*

SPAI – *Sistema Paulista de Ambientes de Inovação*

SPARQL – *Simple Protocol and RDF Query Language*

URL – *Uniform Resource Locator*

URI – *Uniform Resource Identifier*

W3C – *World Wide Web Consortium*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Neste capítulo será abordada a contextualização do trabalho, quais foram as motivações para o seu desenvolvimento, a metodologia de desenvolvimento empregada e sua organização.

1.1 Contextualização

A crescente geração massiva de dados está testando a capacidade das mais avançadas tecnologias de recuperação, armazenamento, tratamento, transformação e análise de informações. As áreas de Gestão e Recuperação da Informação e Apoio à Decisão estão sendo desafiadas pelo volume, variedade e velocidade de uma imensidão de dados semiestruturados e não estruturados de natureza complexa que devem ser encontrados e julgados quanto ao seu valor e veracidade.

O cenário exposto anteriormente evidencia-se pelo fenômeno chamado de Big Data. Desse modo, segundo Manyika et al. (2011) Big Data é definido como um conjunto de dados onde o tamanho está além das capacidades dos bancos de dados típicos, portanto, necessita-se de ferramentas de software específicas para captura, armazenamento, gerenciamento e análise. Além disso, Schroeck et al. (2012) descreve Big Data como sendo uma combinação de volume, variedade, velocidade e veracidade, no qual proporciona para o mercado global atual vantagens competitivas.

Nos ambientes que são caracterizados por Big Data, há a necessidade de empregar novas tecnologias que vão desde a captação dos dados até a persistência, logo, para a obtenção de informação líquida é imprescindível a

utilização de mecanismos de extração de dados adaptados a esse novo contexto, já para a persistência os bancos de dados tradicionais não são pertinentes para o processamento, recuperação e armazenamento, portanto, neste cenário é aconselhável a utilização de bancos não relacionais, como MongoDB, Cassandra, DynamoDB, SimpleDB, etc.

Além disso, segundo Malik et. al. (2011) caracteriza-se os mecanismos de extração de dados ou mecanismos de obtenção de informação líquida como um processo de captação de informações úteis de páginas HTML, onde essas técnicas são semelhantes às utilizadas por motores de busca, mas possuem outro viés que é a estruturação de dados não estruturados e posteriormente sua análise e armazenamento numa base de dados.

Bancos de dados não relacionais mais conhecidos como NoSQL ("*Not Only SQL*") possuem vantagens e desvantagens na utilização em comparação aos bancos relacionais. De acordo com Han et. al. (2011) é vantajoso a sua utilização, pois há o apoio ao armazenamento de grande volume de dados, são fáceis de expandir, tem baixo custo e dispõe de uma leitura e escrita veloz. No entanto, esse tipo de armazenamento falta a possibilidade de transações e relatórios, além de não suportar SQL que é a linguagem de query mais comumente utilizada pela indústria. Dessa forma, é evidente que uma das maneiras benéficas de persistência em cenários de Big Data é a utilização NoSQL.

Com a finalidade de adicionar significado ao que foi extraído pelo mecanismo de captação de informação líquida é necessário combinar o robô de extração com conceitos de Web Semântica, no qual fundamenta-se em aderir significado a páginas Web para a manipulação e processamento de conteúdo por computadores. Portanto, ao realizarmos a extração de um conteúdo desejado num domínio específico obteremos dados semiestruturados e uma busca textual nessas informações em vez de ser sintática será semântica, logo o extraído terá significado e valor que será indispensável para a tomada de decisão. A partir disto, a modelagem Resource Description Framework (RDF) e ontologia aparecem como solução na busca de inserir semântica neste processo.

O RDF é o modelo de descrição de informação recomendado pela W3C (World Wide Web Consortium), utilizado para representar informações disponíveis na Web. Dessa forma, utilizando o SPARQL (SPARQL Protocol and RDF Query Language) linguagem de consulta e protocolo de acesso de dados em RDF há a

possibilidade de obtenção de busca semântica, ou seja, possibilita-se uma busca mais próxima do funcionamento do processo cognitivo do usuário de forma que a extração de dados se torne mais relevante.

Além do RDF, ontologia é outra maneira que aparece como solução para aderir semântica a busca. Segundo Gruber (1992) uma ontologia é uma especificação de uma conceituação. Logo, a aplicação de ontologias no processo de captação de informação líquida tende a possibilitar uma buscar mais inteligente e com dados mais relevantes.

Outrora, outra maneira para se obter uma extração de informação líquida de maneira inteligente é a utilização de computação cognitiva atrelada aos robôs de busca, já que esta área da computação estuda como aderir inteligência a mecanismos computacionais utilizando aprendizado de máquina, probabilidade e estatística e inteligência artificial.

É inegável que atualmente tem crescido substancialmente o volume dos dados, portanto torna-se um grande desafio para as áreas de Ciência da Computação e Ciência Informação buscar formas de gerir, acessar e controlar as informações contidas em cenários de Big Data.

Desse modo, este trabalho tem como objetivo de estabelecer uma Plataforma de Apoio à Inovação que identifica, extrai, classifica, recupera semanticamente e analisa cognitivamente informações advindas de ambientes digitais específicos, onde o espaço informacional é delimitado a Parques Tecnológicos e Centros de Inovação Tecnológica credenciados ao Sistema Paulista de Ambientes de Inovação (SPAI). Para isso, propõe-se a combinação de tecnologias computacionais como os agentes inteligentes de extração de dados, tecnologias informacionais semânticas, como a ontologia e computação cognitiva, sendo que para esta última será empregado o IBM Watson.

Os domínios de aplicação ou espaço informacional do mecanismo de extração de informação proposto nesta pesquisa serão os sites da Fapesp, Inova Marília, Centro Incubador de Empresas de Marília, Parque Tecnológico de Botucatu, entre outros, pois eles possuem diversas informações relevantes sobre ambientes de inovação, parques tecnológicos, incubadoras de empresas, centros e núcleos de inovação tecnológica que são assuntos pertinentes que orientam o estudo.

1.2 Motivação e Objetivos

É de saber mútuo que a colaboração do estado de São Paulo com a regulamentação do SPAI corrobora para uma melhor relação entre governo, empresas e universidades. No entanto, encontra-se uma dificuldade na obtenção de informações referente aos atores de inovação, já que os mesmos possuem objetivos diferentes, as fontes informacionais são descentralizadas e não há padrão estrutural de obtenção de informação sobre inovação.

Além disso, como já foi apresentado a extração de dados está sendo desafiada pelo volume, variedade e velocidade de uma vastidão de dados semiestruturados e não estruturados, logo torna-se o cenário de obtenção de informação líquida referente a atores de inovação mais complexo.

O objetivo deste projeto de pesquisa é o desenvolvimento de uma Plataforma de Apoio à Inovação que extrai, classifica e recupera informações semânticas e cognitivas produzidas pelos atores de inovação (governo, empresas e universidades) em um espaço informacional delimitado a Parques Tecnológicos e Centros de Inovação Tecnológica credenciados ao Sistema Paulista de Ambientes de Inovação (SPAI) que sirva de base para o desenvolvimento de soluções computacionais que possam recuperar informações que poderão ser utilizadas para apoiar a tomada de decisão nos processos de inovação nas organizações. Os objetivos específicos são:

- Identificar técnicas adequadas para identificação, extração, classificação, persistência e recuperação de informações semânticas e cognitivas que compõem o espaço informacional delimitado;
- Estabelecer mecanismos de extração automática de informações (robôs de extração de informações);
- Alterar a ontologia criada por Fusco & Mucheroni & Coneglian (2017) adicionando propriedades nela inexistentes para que seja possível a classificação e persistência semântica;
- Instalar o servidor de triplas denominado FUSEKI para persistência das informações obtidas com os robôs de busca;
- Integrar o IBM Watson à Plataforma de Apoio à Inovação para que seja possível uma análise cognitiva de notícias, editais e eventos;

- Criar uma API que sirva para a integração com aplicações que possuam o objetivo de consumir as informações geradas pela plataforma proposta;
- Desenvolver uma interface Web que apoie o usuário no cadastro, exclusão e recuperação de informações.

1.3 Metodologia de Desenvolvimento do Trabalho

Este trabalho está dividido em nove etapas principais:

1. Levantamento bibliográfico e pesquisa de trabalhos correlatos e tecnologias, sobre os temas de recuperação de informação, robôs de extração, web semântica e computação cognitiva.
2. Estudo do cenário que envolve a classificação de informações e proposta de uma arquitetura informacional semântica e cognitiva (arquitetura informacional da Plataforma de Apoio à Inovação).
3. Desenvolvimento dos robôs de extração para um cenário específico.
    - Nesta etapa será utilizada a linguagem de programação Java e a API JSoup, indicada para extração e manipulação de dados a partir de uma URL. Os robôs de extração devem ser capazes de analisar o conteúdo de uma página WEB e a API Jsoup oferecerá recursos para analisar as TAGs e o conteúdo HTML a partir de uma URL definida.
4. Mapear e classificar as informações extraídas.
    - Nesta etapa o resultado da extração do robô será mapeado na ontologia InovaOnto criada por Fusco (2017), sendo que a biblioteca da Apache chamada JENA possibilitará a aderência de semântica aos resultados obtidos pelo mecanismo de obtenção de informação líquida.
5. Persistir as informações obtidas com os mecanismos de extração.
    - O conteúdo obtido com os robôs de busca depositaremos num banco de dados baseado em triplas chamado FUSEKI. Inclusive, algumas caractéristicas da ontologia (URI) serão armazenadas em um banco de dados relacional para facilitar o

processo de inclusão, remoção e recuperação das informações semânticas.

6. Recuperação de informação.
    - Para a recuperação das informações semânticas será utilizado a linguagem de recuperação semântica SPARQL.
7. Análise Cognitiva.
    - Para a adicionar cognição à plataforma utilizaremos o IBM Watson com os seguintes serviços: “Natural Language Understand” e “Language Translate”.
8. Criação de uma API para integração com outras aplicações.
    - A API será implementada com o auxílio do servidor de aplicação GlassFish e utilizará para comunicação externa Servlets. Além disso, terá funções de inserção, exclusão, listagem de informações e análise cognitiva.
9. Desenvolvimento de uma interface Web para a Plataforma de Apoio à Inovação.
    - A Plataforma de Apoio à Inovação contará com uma interface Web que facilita o usuário utilizar o serviço, além de conter funcionalidades, como: cadastrar e excluir extrações, cadastrar e excluir instâncias estáticas, recuperar informações obtidas e descrição da API criada.

1.4 Organização do Trabalho

Este trabalho está organizado da seguinte maneira. Primeiramente, no capítulo 2 é elucidado temas contidos no estado da arte referente a fundamentação teórica, como: Big Data, Web Semântica, Metadados, padrão de metadados Dublin Core, RDF, Ontologia, a biblioteca Jsoup, a ferramenta da Apache chamada Jena, o servidor FUSEKI, a linguagem de consulta SPARQL e IBM Watson.

Ademais, no capítulo 3 é demostrado os trabalhos correlatos, onde são abordados aqueles que de certa forma contribuem para este trabalho com a construção de uma base referencial sólida.

No capítulo 4, é mostrado a arquitetura informacional que engloba este projeto, qual o fluxograma da Plataforma de Apoio à Inovação, as principais funcionalidades e suas telas.

Além disso, no capítulo 5 é mostrado como foi implementada a Plataforma de Apoio à Inovação e suas funcionalidades que vão desde a extração de metadados e dados em ambientes informacionais digitais até a recuperação desse conteúdo do servidor do FUSEKI.

Já no capítulo 6, é demonstrado os resultados, considerações finais, contribuições, limitações e produção bibliográfica deste Trabalho de Conclusão de Curso.

Capítulo 2

CONCEITOS, TECNOLOGIAS E FERRAMENTAS

Para que seja desenvolvida a Plataforma de Apoio à Inovação necessita-se de um embasamento teórico, portanto, este capítulo possui o objetivo de abordar conceitos, tecnologias e ferramentas encontrados no estado da arte, como: Big Data, Web Semântica, Metadados, Dublin Core, RDF, Ontologia, Jsoup, Apache Jena, Apache FUSEKI, SPARQL e IBM Watson.

2.1 Big data

Atualmente há a uma crescente onda de dados que está pondo em cheque as tecnologias de recuperação, armazenamento, tratamento, transformação e análise de informações. Desse modo, as áreas de Gestão e Recuperação da Informação estão sendo instigadas pela variedade, velocidade e volume que estes dados se apresentam.

Uma pesquisa realizada por pesquisadores da Escola de Gerenciamento de Informação e Sistemas da Universidade de Berkeley, estima-se que a humanidade acumulou cerca de 12 *exabytes* (12 x 10^18) até a década de 90 e somente no ano de 2002 os estudiosos constataram que os seres humanos produziram cerca de 5 *exabytes*. Desse modo, para que seja possível persistir toda essa informação em papel é necessário a construção de 37 mil novas bibliotecas com o tamanho equivalentes ao tamanho da Biblioteca do Congresso Americano, (Floridi, 2010).

O cenário descrito anteriormente destaca-se como Big Data. Dessa forma, de acordo com Bayer e Laney (2012, p. 1-9, apud Mauro, 2014) define que este cenário possui como característica um grande volume, velocidade e variedade de recursos informacionais, portanto exige-se formas inovadoras para o processamento da informação e assim tomar uma determinada decisão.

Além disso, Schroeck et al. (2012) elucida que a fenômeno definido como Big Data possui uma exorbitante quantidade de dados informacionais, sendo que estes possuem peculiaridades que devem ser analisadas, como: volume, variedade, velocidade, veracidade. Desse modo, quando bem interpretados garantem as organizações vantagens competitivas no mercado digitalizado de hoje.

Ademais, Ward e Barker (2013) informa que a análise e armazenamento de uma exacerbada quantidade de dados informacionais necessita técnicas específicas para análise, como *MapReduce* e Aprendizagem de Máquina, já para armazenamento é comumente utilizado bancos não relacionais, ou seja, NoSQL.

Levando em consideração o cenário exposto, conclui-se que o conceito de Big Data pode ser elucidado conforme cinco Vs, no qual são: Volume, Velocidade, Variedade, Veracidade e Valor. A seguir encontra-se uma representação imagética do conceito de Big Data e suas peculiaridades citadas anteriormente:

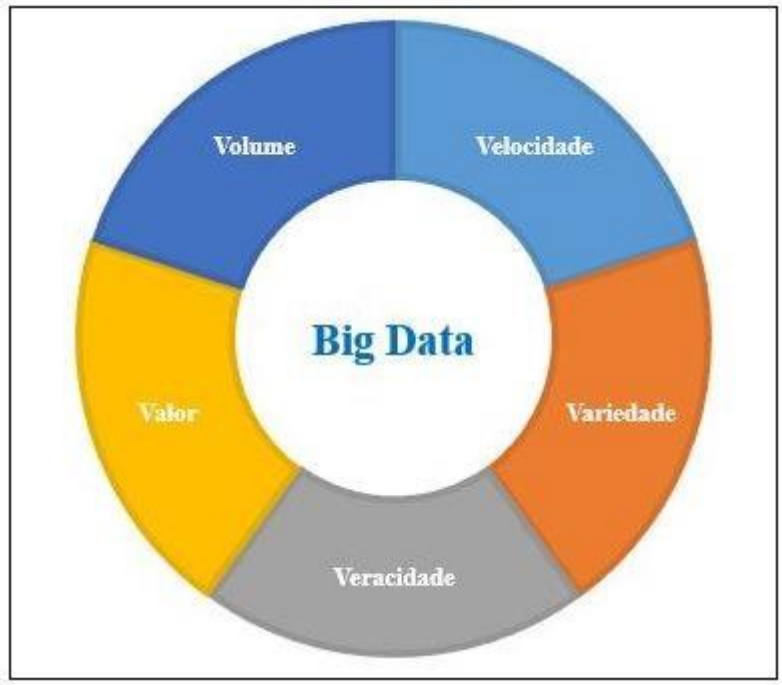


Figura 2.1: Big Data

Dessa maneira, as peculiaridades do fenômeno denominado Big Data, podem ser descritas da seguinte forma:

1. **Volume:** Um estudo realizado por um grupo de pesquisadores da Escola de Gerenciamento de Informação e Sistemas da Universidade de Berkeley constatou-se que a quantidade de dados digitais gerados entre 2006 a 2010 cresceu de 166 exabytes para 988 exabytes, (Floridi, 2010). Observa-se com esta informação que os sistemas tradicionais de armazenamento não são páreos para esta quantidade, ou seja, há a necessidade da implantação de novas tecnologias de armazenamento.
2. **Velocidade:** É imprescindível a análise da velocidade em que as informações são criadas, manipuladas e descartadas nos tempos de hoje, pois segundo Hsu et al. (2010) a velocidade que se obtêm uma determinada informação dita uma vantagem competitiva em relação aos seus concorrentes, por exemplo, ao utilizar um cartão de crédito, caso o mesmo demore mais do que alguns instantes para a efetivação da compra, logo os usuários tendem em utilizar outro método de pagamento. Portanto, a operadora consequentemente devido a falha de transmissão de seu serviço de prestação pode acabar perdendo clientes.
3. **Variedade:** Atualmente a rede mundial de computadores é utilizada com diversos objetivos e por adversos usuários, sendo que os usuários podem ser pessoas que acessam redes sociais, páginas de gastronomia, esportes, trabalhos acadêmicos, compras etc. e máquinas que possuem o objetivo de compartilhar informações, como sistemas meteorológicos, câmeras de segurança, sistemas bancários, sistemas hospitalares e outros. Dessa forma, cada tipo de informação transitada por esses espaços informacionais acessados por diversos usuários possui suas peculiaridades, onde em Big Data podem ser classificadas essas informações como estruturadas, semiestruturadas e não estruturadas.
4. **Veracidade:** Kakhani (2013, apud Coneglian, 2014, p.26), define que os dados presentes neste universo possuem adversas naturezas, portanto para um determinado estudo há a necessidade que eles sejam verdadeiros para que não transmita informações equivocadas.

5. **Valor:** De acordo com os dados coletados há a possibilidade da realização de predições, por exemplo, com base nas informações obtidas de todas as temporadas dum ano referente a agricultura é possível com base em modelos probabilísticos, como as cadeias ocultas de Markov, falar qual a probabilidade de que haja seca numa área e em um determinado período do ano.

Conforme foi explicitado anteriormente o fenômeno Big Data quando analisado como um todo é possível realizar predições baseado nos dados coletados por meio dos sistemas, logo a seguir apresenta-se casos de sucesso da utilização deste fenômeno e do impacto que o mesmo tende na sociedade.

Segundo Anderson e Raine (2012, apud Moura e Amorim, 2014), em um estudo realizado por centenas de pesquisadores e especialistas retrata-se os impactos positivos e negativos que a era da Big Data ocasionará para as empresas, sendo que 53% dos entrevistados possuem um posicionamento positivo relacionado a Big Data, já 39% tem um posicionamento negativo. Logo, um dos impactos positivos presentes é que surgirá novos empregos para as pessoas e a criação de um novo cargo o "*Data Scientist*", este cargo de cientistas de dados é direcionado para profissionais formados em Ciência da Computação e Matemática e sua função é analisar e gerenciar este grande fluxo de dados.

Em prol de formar profissionais capacitados para este ramo em 2012 a IBM criou um programa chamado *Big Data University*, no qual possui a função de oferecer oportunidades de aprendizado para estudantes de graduação e pós-graduação na área com o intuito de capacitar estes alunos em tecnologias, como *Hadoop* e conceitos de Big Data, (IBM, 2012).

Dessa forma, WEF (2012, apud Moura e Amorim, 2014) elucida que a quantidade de dados informacionais produzidos pelos usuários de celulares tem despertado a atenção de líderes políticos e empresários, já que o crescimento do tráfego de dados móveis nos países emergentes ultrapassa 100% ao ano. Desse modo, com esses dados é possível prever surtos endêmicos, necessidades e comportamentos da população, construir soluções centradas nos usuários para fornecer melhores serviços de saúde, transporte, educação, serviços financeiros etc.

Além disso, WEF (2012, apud Moura e Amorim, 2014) informa que o fenômeno Big Data surge efeito em áreas como Serviços Financeiros e Agricultura,

sendo que na primeira área dados recolhidos a partir de pagamentos fornece uma visão sobre gastos e hábitos de poupança em determinada região, além de que o histórico de créditos de uma pessoa pode influenciar torna-la uma possível candidata a empréstimos por parte de serviços financeiros. Não menos importante, a compra de produtos agrícolas, aquisição de insumos e subsídios ajudam o governo a prever a tendência da produção de alimentos e incentivos, onde esse conhecimento pode ser utilizado para garantir a qualidade e armazenamento de determinadas culturas e reduzir a quantidade de resíduos e desperdícios.

Conclui-se que a Big Data é um fenômeno que surgiu com vinda da globalização, já que neste período aumentou-se a produção de dados e este fenômeno possui algumas características, como: volume, velocidade, variedade, veracidade e valor. Dessa forma, tais peculiaridades quando estudas e analisadas fornecem tomadas de decisões mais precisas e vantagens competitivas no mercado global atual.

2.2 Web Semântica

A atual Web em seus primórdios possuía como característica suas páginas serem desenvolvidas por programadores e engenheiros com a finalidade de compartilhamento de informação. Dessa forma, posteriormente popularizou-se ferramentas que proporcionaram a familiarizados ou não com o conhecimento de programação o desenvolvimento de páginas da Internet, (Breitman K., 2005).

Com o passar dos anos o cenário da Web tende a crescer cada vez mais, no entanto a mesma possui a peculiaridade que em suas páginas as informações contidas são destinadas a leitura de seres humanos e não para a manipulação, processamento e interpretação da informação por máquinas e seus *softwares*.

Segundo Breitman (2005) em seu livro intitulado “Web Semântica: A Internet do Futuro”, a mesma denomina a Internet atual como Web Sintática, já que os computadores fazem apenas a apresentação da informação, porém os seres humanos são responsáveis pelo processo de interpretação. Ademais, Marcondes (2012) acrescenta que a Web atual trabalha com padrões de *forma*, ou seja, uma

pesquisa por um determinado conteúdo é buscada pelos mecanismos de consulta de forma idêntica ao fornecido por parâmetro.

A seguir, apresenta-se uma imagem elucidando o conceito de Web Sintática, no qual a função da máquina é somente de intermédio na apresentação das informações entre os autores do conteúdo e os leitores, logo, a interpretação da informação está a cargo do ser humano.

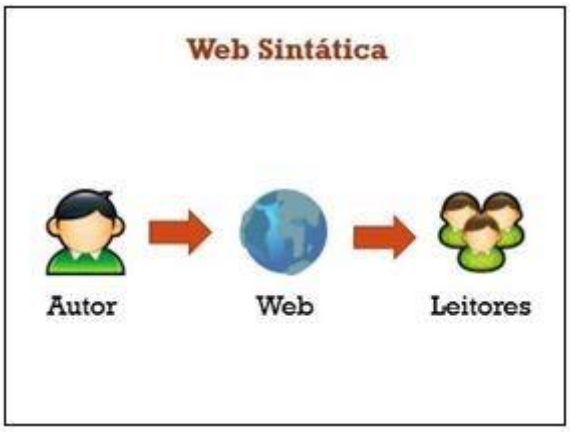


Figura 2.2: Web Sintática

A autora Breitman (2005) denomina alguns problemas que os mecanismos de buscas vivenciaram na época da escrita de seu livro, como ferramentas do tipo Google, Yahoo, Bing etc.:

- Grande número de páginas encontradas, porém com pouca precisão: ou seja, ao realizar uma busca por um determinado assunto a resposta terá pouca utilidade se recuperar as páginas relevantes para minha pesquisa, mas também trazer 39.857 páginas de interesse médio ou pouco relevantes. Dessa forma, a pessoa que está realizando a pesquisa numa determinada ferramenta de busca demandará um esforço para encontrar aquilo que é realmente necessário para ela;
- Resultados são muito sensíveis ao vocabulário: isto é, a ordem das palavras que é realizado a busca influência no resultado da consulta e documentos importantes utilizam uma terminologia distinta da nossa falada;

- Resultados são páginas individuais: quer dizer, ao realizar uma pesquisa em mecanismos de busca os resultados obtidos como resposta são diversos e alguns são pertencentes ao mesmo site.

A solução para as máquinas e seus softwares terem a capacidade de conseguir manipular, processar e interpretar informações advindas de páginas da Internet chama-se Web Semântica, onde conceitua-se em aderir semântica nos dados da Web para que os *softwares* sejam capazes de manipulá-los conforme o ser humano deseja.

A palavra semântica é de origem grega *semainô* (significar), derivado da palavra *sema* (sinal), que corresponde a sentido. Ou seja, tudo que se refere a um sentido de um sinal de comunicação e tudo que se diz respeito as palavras (Guirald, 1980, apud Pickler, 2007).

Desse modo, conforme Pickler (2007) como a palavra semântica se encarrega do estudo das palavras, portanto Web Semântica possui a função de aderir semântica ao conteúdo da Web, já que essa semântica servirá para estabelecer o sentido de um termo no contexto de um determinado documento.

A Web Semântica é um conceito que foi desenvolvido pelo criador e idealizador da *World Wide Web* (WWW), Tim Berners-Lee, no qual tal conceito infere características na Web para torna-la mais inteligente e intuitiva para atender de uma maneira mais específica as necessidades de um usuário.

No ano de 2001, Berners-Lee et al. (2001) publicaram na revista Scientific American o artigo "*The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*", onde traduz-se para o português do Brasil como "Web Semântica: uma nova forma de conteúdo para Web que tem significado para computadores e vai iniciar uma revolução de possibilidades".

Neste artigo Berners-Lee et al. (2001), o autor expõe que a Web Semântica apresenta significado aos computares e desencadeará uma onda de oportunidades. Inclusive, o mesmo cita que a Web Semântica não é uma Web separada, mas sim uma extensão da atual, onde a informação tem significado bem preciso e isso propicia os computadores e as pessoas trabalharem de forma colaborativa.

Além disso, segundo Brascher (2007) a Web Semântica é uma plataforma acessível e universal para permitir que os dados possam ser compartilhados e

processados por pessoas e por ferramentas automáticas, ou seja mecanismos computacionais capazes de buscar, filtrar e preparar a informação para os seres humanos.

De acordo com Berners-Lee et al. (2001), as aplicações da Web Semântica podem ser diversas, logo os autores explicam didaticamente as implicações da mesma na vida das pessoas, portanto a seguir observa-se dois cenários:

1. Uma pessoa precisa marcar uma consulta com um médico, logo ela notifica o computador e lhe informa algumas limitações. O computador navega pela Internet procurando médicos que estejam perto da residência do usuário, que sejam conveniados ao seu plano de saúde e possuam uma boa reputação. Desse modo, de uma maneira inteligente a máquina compara as agendas de consulta do médico e seu horário de atendimento com a agenda da pessoa que está solicitando o serviço, assim oferece opções de atendimento. A pessoa que está consultando a máquina deve se preocupar somente com o horário que lhe convém.
2. Uma pessoa está em sua casa e num determinado momento seu telefone toca, automaticamente o volume de todos seus eletrodomésticos abaixa. Tal feito é possível, pois a mesma cadastra num dispositivo todos os seus eletrodomésticos, como TV, DVD, computador, babá eletrônica, etc., ou seja, todos os dispositivos que emitem áudio. Desse modo, quando seu telefone tocar, automaticamente todos os dispositivos com controle de volume diminuem para que a pessoa possa falar no telefone tranquilamente.

Portanto, do modo que foi explicitado anteriormente observa-se como a Web Semântica facilita o cotidiano dos seres humanos. Além disso, introduz no dia a dia das pessoas a colaboração das máquinas para a resolução de problemas cotidianos que demandariam um determinado esforço desde físico até temporal.

Conforme Dias e Santos (2013), a capacidade da Web Semântica só chegará em seu apogeu quando forem desenvolvidos *softwares* que tenham a capacidade de obter informação de diferentes fontes informacionais e compartilhar tais resultados com outros programas. Consequentemente, a eficiência dos agentes aumentará quando toda Web esteja estruturada de forma semântica, pois os mecanismos de

extração captarão as necessidades do usuário, pesquisarão e disponibilizarão para consulta.

Na imagem a seguir elucida-se o conceito explicitado anteriormente, em que a mesma é responsável pelo compartilhamento de informações entre pessoas e ferramentas e esses dados podem ser processados, interpretados e manipulados pelos mesmos.

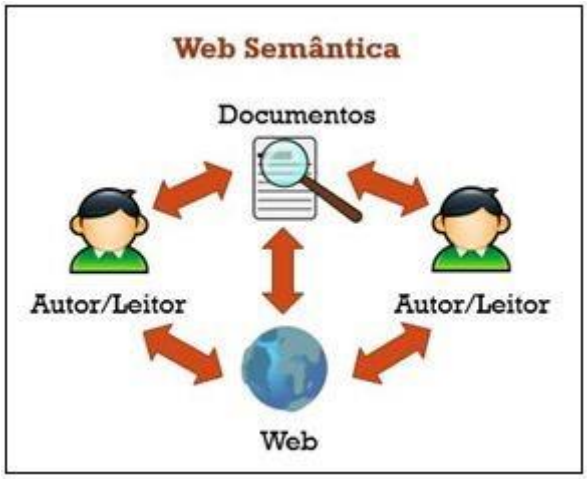


Figura 2.3: Web Semântica

A partir disso, há várias formas de se fazer a Web Semântica se tornar real, logo na imagem a seguir há o modelo de camadas estipulado pelo W3C:

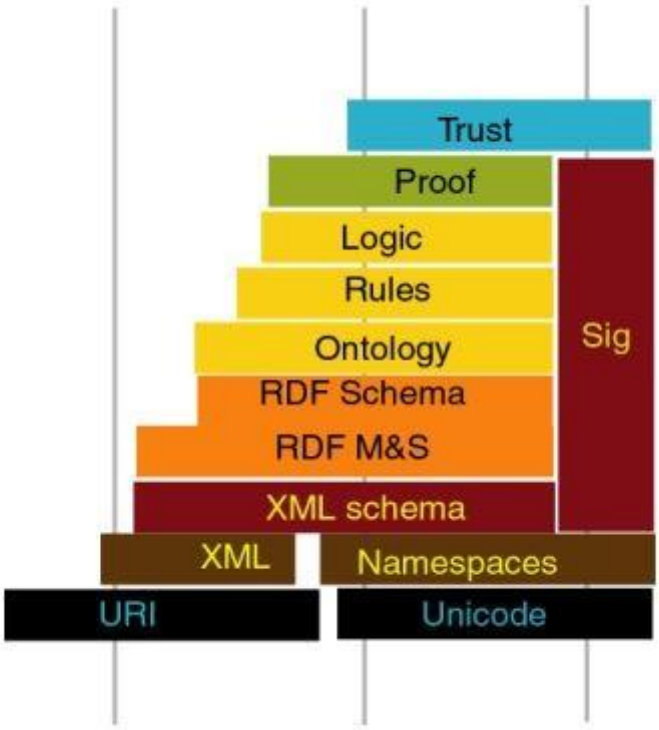


Figura 2.4: Estrutura da Web Semântica – Fonte: (W3C, 2014h, apud Coneglian, 2014, p.35-36)

Cada camada a seguir é descrita segundo Coneglian (2014, p. 35-36):

- URI (Uniform Resource Identifier – Identificador de Recursos Uniforme): conjunto de caracteres para a identificação de um recurso (W3C, 2014b, apud Coneglian, 2014, p. 35-36);
- Unicode: define um conjunto e padrão universal de codificação (UNICODE, 2008 apud Coneglian, 2014, p.35-36);
- XML (Extensible Markup Language – Linguagem de Marcação Extensível): é um sistema de representação de informação estruturada (W3C, 2014c, apud Coneglian, 2014, p.35-36);
- 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, apud Coneglian, 2014, p.35-36);
- RDF M&S: um modelo para intercâmbio de dados na web, e tem características que facilitam a fusão de dados (W3C, 2014e, apud Coneglian, 2014, p.35-36);
- RDF Schema: um vocabulário para fazer a modelagem de dados de RDF (W3C, 2014f, apud Coneglian, 2014, p.35-36);
- Ontology: é um modelo de dado que representa um conjunto de conhecimento e o relacionamento entre eles dentro de uma base informacional;
- Rules: nela é feita a conversão das informações que estão dentro de um documento para outro, criando regras de inferência (PRADO, 2004, apud Coneglian, 2014, p.35-36);
- 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.

2.3 Metadados

Segundo Breitman (2005), metadados são dados que descrevem dados, ou seja, dados sobre dados. Logo, eles propiciam a descrição de elementos de documentos, como: data, autor, editora, etc. Ademais, Morato e Moraes (2014) salienta que que os metadados também são uma forma de descrever recursos eletrônicos dispostos na Web e apresentam uma semântica padronizada que a descrição de recursos eletrônicos de maneira bibliográfica.

Além disso, Breitman (2005) informa que para a *International Federation of Library Associations* (IFLA) o termo se refere a qualquer informação utilizada para a identificação, descrição e localização de recursos. Ademais, Breitman (2005) indica que para a *World Web Consortium* (W3C) o termo está mais relacionado com a Web Semântica, já que a mesma determina que metadados são informações que são compreendidas por máquinas.

> Um dos maiores objetivos do uso de metadados no contexto da Web é permitir não só descrever documentos eletrônicos e informações em geral, possibilitando sua avaliação de relevância por usuários humanos, mas também permitir agenciar computadores e programas especiais, robôs e agentes de software, para que eles compreendam os metadados associados a documentos e possam então recuperá-los, avaliar sua relevância e manipulá-los com mais eficiência (MARCONDES, 2005, p. 96).

Além disso, a seguir observa-se uma tabela de Gilliland-Swetland (2000, apud Breitman, 2005), no qual demonstra os tipos e funções dos metadados.

Tabela 2.1: Tipos e funções dos metadados – Fonte: Gilliland-Swetland (2000, apud Breitman, 2005)

Tipo	Definição	Exemplos
Administrativo	Metadados utilizados na gerência e na administração de recurso de informação	Aquisição de informação Registro de direitos e reprodução Documentação dos requisitos legais de acesso Informação de localização

		Critérios de seleção para a digitalização Controle de versão
Descritivo	Metadados utilizados para descrever e identificar recursos de informação	Registros de catalogação Auxílio para a procura de informação Indexes especializados Utilização de hiperlinks entre recursos Anotações
Preservação	Metadados relacionados ao gerenciamento dos recursos de informação	Documentação sobre a condição física dos recursos Documentação sobre as ações tomadas de modo a preservar as versões físicas e digitais dos recursos, e.g., atualização e migração
Técnica	Metadados relacionados a funcionalidades do sistema e como seus metadados se comportam	Documentação sobre hardware e software Informação relativa a digitação, e. g., formatos, compressão, rotinas de escalonamento Registro de tempo de resposta do sistema Autenticação de dados, e.g., senhas e criptografia
Utilização	Metadados relacionados ao nível e o tipo de utilização dos recursos	Registro de Exibição Registro do uso e dos usuários dos recursos Reutilização de conteúdo e multiversionamento de informação

Portanto, no cenário de bibliotecas o que se denominava registro bibliográfico com o passar do tempo foi adaptado para o contexto da era digital e foi nomeado de metadados, onde possui a função de descrever recursos digitais depositados ou armazenados na internet (Morato; Moraes, 2014).

2.4 Dublin Core

O padrão de metadados Dublin Core segundo Souza (2000), é definido como um conjunto de elementos descritivos que possuem a finalidade de descrever recursos digitais presentes na Web.

Dessa forma, de acordo com Breitman (2005), o padrão Dublin Core surgiu durante uma das primeiras conferências sobre Web em 1994 situada em uma cidade no Estados Unidos da América chamada Dublin, foi apontado uma necessidade de padronizar semanticamente os recursos digitais presentes na Internet. Portanto, nesta conferência idealizaram que a descrição dos recursos digitais fosse organizada semelhantemente a um “cartão virtual”, no qual nele é descrito os elementos que compõem o recurso.

Segundo Breitman (2005), os quinze elementos descritivos do padrão de metadados Dublin Core são:

Tabela 2.2: Elementos que compõem o padrão Dublin Core – Fonte: Breitman (2005)

Asssunto (subject)	**O tópico abordado pelo trabalho**
Título (title)	Nome do objeto
Criador (creator)	Pessoa(s) responsável(eis) pelo conteúdo intelectual do objeto
Descrição (description)	Descrição do conteúdo do objeto
Editor (publisher)	Agente ou agência responsável por disponibilizar o objeto
Outro agente (contributor)	Pessoa(s) que fez (fizeram) contribuições significativas para o objeto
Data (date)	Data de publicação
Tipo de Objeto (type)	Gênero do objeto, se ficção, novela, poema ou dicionário
Formato (format)	Manifestação física do objeto. Exemplos são arquivos executáveis, do tipo texto ou PDF
Identificador (identifier)	Cadeia ou número utilizado para identificar unicamente aquele objeto
Relacionamento(relation)	Relacionamento com outros objetos
Fonte (source)	Outros objetos, eletrônicos ou físico, dos quais este foi derivado (caso seja aplicável)
Linguagem (language)	Linguagem do conteúdo intelectual
Cobertura (coverage)	Localizações espaciais e durações temporais características do objeto
Direitos (rights)	Informações sobre os direitos acerca do objeto

Além disso, de acordo com Morato & Moraes (2014) os elementos do Dublin Core podem ser organizados em três grupos, sendo eles:

Tabela 2.3: Características das informações que contemplam os elementos do Dublin Core – Fonte: Morato & Moraes (2014)

Relacionados com o conteúdo	**Relacionados com a propriedade intelectual do recurso**	**Relacionados com características formais do recurso**
Title	Creator	Date
Subject	Contributor	Format
Relation	Publisher	Identifier
Source	Rights	Language
Coverage		
Type		

Portanto, conclui-se que padrão Dublin Core é um modelo bem simples que cumpre a sua função de descrever recursos digitais presentes na Web, sendo que esta simplicidade contribui para a disseminação do mesmo em larga escala.

2.5 RDF

O *Resource Description Framework* (RDF) é um modelo de descrição recomendado pelo W3C, onde é utilizado para representar informações disponíveis na Internet. Conforme Breitman (2005), o RDF possibilita que seja criado relacionamentos entre itens presentes na Web e como é uma linguagem declarativa que fornece uma maneira padronizada de utilizar XML é possível representar metadados no formato de sentenças sobre propriedade.

Além disso, segundo Ramalho (2006) o modelo de especificação de sintaxe do RDF foi proposto em 1999 pelo W3C, no qual surgiu com o intuito de aumentar a interação e comunicação no ambiente Web. Ademais, Ramalho (2006) acrescenta que o padrão o RDF possui uma gama altíssima de aplicações sendo possível fazer qualquer declaração de qualquer tipo de objeto, contanto que o mesmo possua um endereço URI.

De acordo com Breitman (2005), um dos objetivos do RDF é aderir semântica ao ambiente Web e tornar possível o acesso de seus recursos por parte das máquinas. Além disso, a autora elucida alguns exemplos de utilização de RDF, como: descrição de conteúdo para mecanismos de extração de informação, especificação de informações sobre páginas da Internet, descrição de conteúdo e classificação de figuras e especificação de propriedades para itens de compra, tais como preços e suas disponibilidades.

Desse modo, observa-se que o RDF possui o objetivo de prover um mecanismo de descrição de documentos que não esteja ligado a nenhum domínio de conhecimento específico, inclusive, o mesmo possibilita que haja uma interoperabilidade entre aplicações através de uma permuta de informações estruturadas e a automação de processos na Web (Dias; Santos, 2003).

Santarém Segundo e Vidotti (2011) um RDF é constituído de três objetos básicos, sendo eles: recursos, propriedades e declarações. Portanto, um recurso é uma informação que pode ser identificado por uma URI (*Uniform Resource Identifi*er) e as propriedades são informações que representam as peculiaridades do recurso e o relacionamento entre eles, já as declarações é a constituição da informação completa.

Além disso, conforme Klyne et. al. (2004, apud Ramalho, 2006), os princípios fundamentais do padrão RDF é conforme a tripla “*subject*, *predicate* e *object*”, logo, na tradução para português do Brasil é respectivamente “sujeito, predicado e objeto”. Desse modo, “sujeito” refere-se ao recurso que uma sentença está se referindo, já “predicado” descreve uma característica, propriedade ou relacionamento usado para descrever este recurso e o “objeto” é o valor de uma determinada característica do recurso.

Dessa forma, Ramalho (2006) conclui que em RDF toda sentença é chamada de declaração e é formada por um “sujeito” que está relacionado a um “predicado” que possui um valor indicado a partir de um “objeto”. É comumente encontrado na bibliografia a utilização da tripla “recurso”, “propriedade” e “valor” no lugar de “sujeito”, “predicado” e “objeto” respectivamente. Logo, neste trabalho adotaremos a triplo “recurso”, “propriedade”, “valor” como padrão.

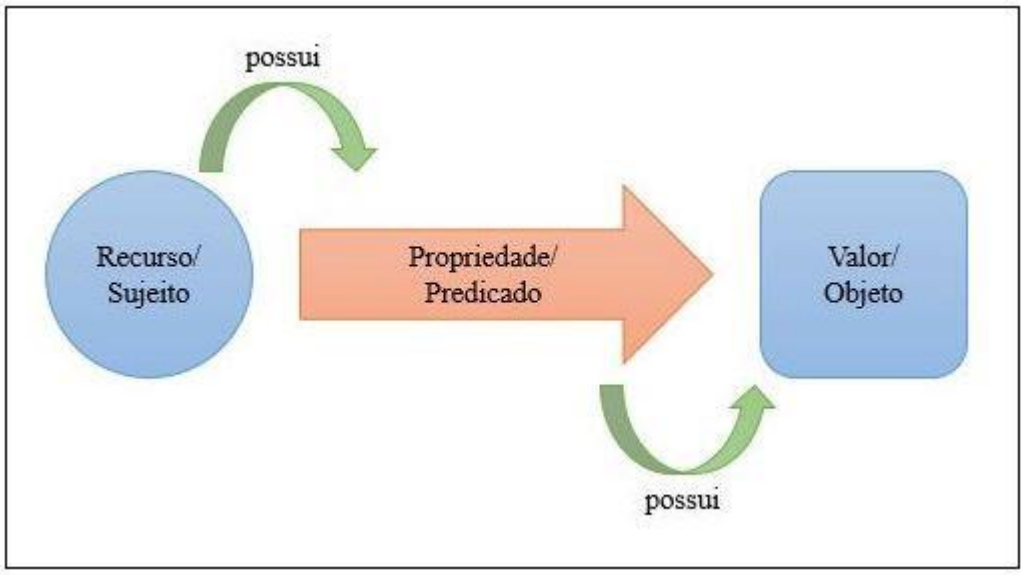


Figura 2.5: Tripla do RDF

Como explicado anteriormente o RDF utiliza uma URI para identificar um determinado recurso na Web e propriedades para descrever esse recurso, logo a seguir encontra-se um exemplo:

```
<?xml version="1.0"?>
 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
   <rdf:Description rdf:about="https://github.com/thiagogcosta/MLP-Backpropagation">
     <dc:title>Repositório do trabalho MLP-Backpropagation</dc:title>
     <dc:creator>Thiago Costa</dc:creator>
     <dc:date>30 de maio de 2017</dc:date>
   </rdf:Description>
 </rdf:RDF>
```

Figura 2.6: Estrutura do RDF/XML

Conforme a figura demonstrada anteriormente, observa-se algumas características presentes na estrutura do RDF, sendo elas responsáveis por distinguir quem é o recurso, propriedade e o valor no código, logo a seguir observa-se essa distinção:

- Primeiramente, “https://github.com/thiagogcosta/MLP-Backpropagation” é um **recurso**, pois em RDF um recurso é tudo aquilo que possui uma URI, (Breitman, 2005);
- Além do mais, “<dc:title>”, “<dc:creator>”, “<dc:date>” são **propriedades** segundo Breitman (2005), pois são recursos que possuem um nome para caracterizar outros recursos, como por exemplo os citados respectivamente: título, criador e data. Outrora, o prefixo “dc” contido em “<dc:title>” é a sigla de Dublin Core, ou seja, na escrita deste RDF está sendo utilizado o padrão de metadados Dublin Core;
- Ademais, as cadeias de caracteres do RDF chamadas de “Repositório do trabalho MLP-Backpropagation”, “Thiago Costa” e “30 de maio de 2017” são **valores** das seguintes propriedades, respectivamente: título, criador e data.

Inclusive, Breitman (2005) salienta que RDF pode ser escrito em XML/RDF e isso é feito para que haja uma interoperabilidade entre diferentes máquinas e sistemas operacionais. Além de que, a arquitetura proposta por Tim Berners Lee para a Web Semântica é baseada em sobreposição de camadas em cima do XML, pois isso garante que haja uma maior expressividade por parte da camada que está sobrepondo e assegura que os computadores que estão processando conseguem minimamente processar a porção XML do arquivo mesmo que não entendam as camadas superiores.

Ademais, um RDF pode ser representado de outras maneiras, sendo elas: através de uma tabela ou um grafo, logo a seguir demonstra-se essas informações, no qual as mesmas foram retiradas da figura 2.7.

Tabela 2.4: Representação do RDF por meio de Tabela

Sujeito (recurso)	Predicado (propriedade)	Objeto (valor)
https://github.com/thiagogcosta/MLP-Backpropagation	http://purl.org/dc/elements/1.1/title	Repositório do trabalho MLP-Backpropagation
https://github.com/thiagogcosta/MLP-Backpropagation	http://purl.org/dc/elements/1.1/creator	Thiago Costa
https://github.com/thiagogcosta/MLP-Backpropagation	http://purl.org/dc/elements/1.1/date	30 de Maio de 2017

Ou grafo:

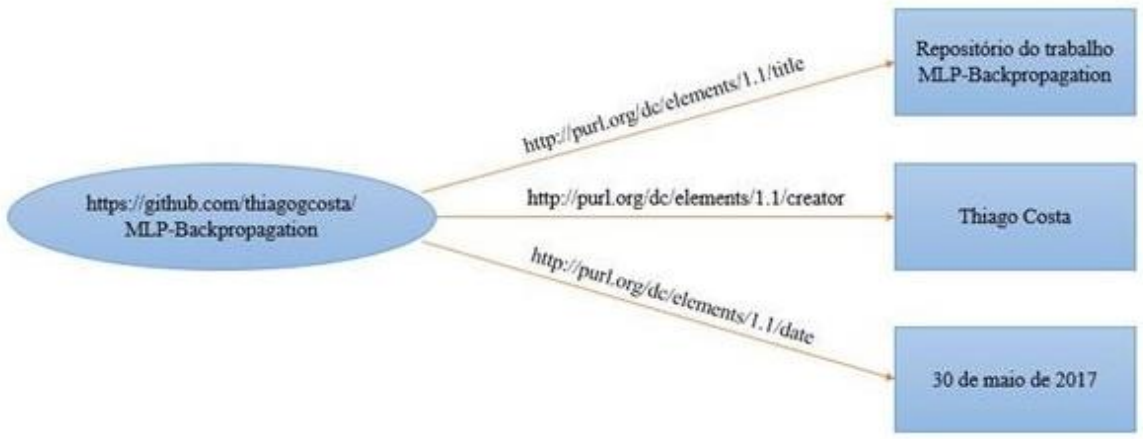


Figura 2.7: Grafo RDF

A função exercida pelo esquema RDF na criação, descrição e combinação de classes de tipos de recurso e propriedades e na atribuição de valores e restrições entre relacionamentos, define que o esquema pode ser utilizado em conjunto com vocabulários descritivos, como o Dublin Core, Dias e Santos (2003).

Conclui-se que o RDF possui a função de aderir semântica ao ambiente Web e consequentemente proporcionar as máquinas acesso dos recursos desse ambiente a elas, sendo que esta arquitetura de descrição de documentos foi criada por Tim Berners Lee e pode ser aplicada em diferentes ocasiões, essas que vão desde a descrição de conteúdo para agentes de extração de informação até especificação de páginas da Internet, imagens, vídeos etc.

2.6 Ontologia

A ontologia, é definida por Silva (2003) como a parte da ciência que estuda o ser e seus relacionamentos, o uso de ontologias é fundamental no processo de desenvolvimento do conhecimento conceitual informacional, dos robôs de busca semântica, sendo aplicada na Ciência da Computação e na Ciência da Informação para condicionar uma busca de maneira mais inteligente e mais próxima do funcionamento do processo cognitivo do usuário de forma que a extração de dados se torne muito mais relevante.

Além disso, de acordo com Gruber (1993) ontologia é uma especificação explícita de uma conceituação de um 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, pois é uma técnica de organização de conceitos e relacionamentos.

Na Ciência da Computação, Guarino (1998) diz que a ontologia é uma teoria lógica que representa um vocabulário pretendido, ou seja, é uma contextualização de algo particular existente no mundo. Neste sentido observa-se que com uma ontologia você consegue definir contextos e domínios particulares do mundo.

Há diversas linguagens construídas para a implementação de ontologias, com destaque para a Web Ontology Language (OWL), linguagem recomendada pela W3C.

Segundo a W3C a OWL é uma linguagem da Web Semântica projetada para representar o conhecimento rico e complexo sobre coisas, grupos de coisas e relações entre as coisas. OWL é uma linguagem baseada em lógica computacional que pode ser explorada por programas de computador, por exemplo, para verificar a

consistência desse conhecimento ou para tornar explícito o conhecimento implícito. Os documentos escritos em OWL, conhecidos como ontologias, podem ser publicados na Web ou referenciar outras ontologias.

2.7 Jsoup

Segundo Hedley (2016), Jsoup é uma biblioteca para a linguagem de programação Java que possibilita trabalhar com HTML, onde há a possibilidade de extrair e manipular dados, analise da estrutura HTML de uma URL ou sequência de caracteres, manipulação de elementos HTML e texto, além disso, possibilita-se encontrar elementos estruturais através de passagem de DOM (Document Object Model) e CSS. Ademais, Jsoup implementa a especificação WHATWG HTML5 que proporciona trabalhar com DOM da mesma forma que os navegadores contemporâneos operam, ou seja, proporciona uma árvore de objetos mais sensata possível ao usuário da ferramenta.

2.8 Apache Jena

Apache Jena é um framework de código aberto criado pela Apache Software Fundation para a criação de aplicações contendo conceitos de Web Semântica e Dados Conectados. O Jena possui uma série de ferramentas, portanto segue algumas ferramentas:

- API de RDF: ferramenta para a criação e leitura de RDF;
- ARQ (SPARQL): mecanismo que proporciona a consulta em RDF utilizando uma arquitetura denominada ARQ;
- API de OWL: ferramenta que possibilita o trabalho com modelos ontológicos descritos em OWL (Web Ontology Language).

2.9 Apache Fuseki

O apache FUSEKI é um servidor baseado em triplas, ou seja, é um servidor SPARQL. Sendo que, nele há a possibilidade de inserir ontologias e rdfs em sua base de dados para a persistência de informações, o serviço proporciona um painel Web completo para a inserção, alteração e exclusão de base de dados.

Além disso, o FUSEKI proporciona ao usuário uma integração com serviços externos, pois ele disponibiliza funcionalidades através de requisição HTTP de carregamento, atualização e exclusão de arquivos (RDF, OWL, etc.).

2.10 SPARQL

SPARQL (Simple Protocol and RDF Query Language) é uma línguagem de consulta em dados armazenados em RDF. Portanto, utiliza-se SPARQL para a consulta de informações contidas numa ontologia.

Dessa forma, segundo Speroni (2014) esta linguagem de consulta permite recuperar valores de dados estruturados e semiestruturados, explorar relações contidas numa ontologia e realizar uniões complexas de um conjunto massivo de dados numa única consulta.

Ademais, Speroni (2014) elucida que existe quatro formas de consulta via SPARQL:

- **SELECT:** Retorna todas ou um subconjunto das variáveis pesquisadas;
- **CONSTRUCT:** Retorna um RDF com variáveis substituídas por um conjunto de modelos de triplas;
- **ASK:** Informa através de um indicador se um padrão de consulta foi achado;
- **DESCRIBE:** Retorna um RDF descrevendo os recursos filtrados.

As consultas SPARQL são realizadas através de triplas RDF (sujeito, predicado, objeto), portanto a seguir na figura 2.8 há um exemplo de como se faz uma busca numa ontologia ou uma base de dados RDF utilizando a linguagem de

consulta SPARQL, onde como resultado trará todos os sujeitos, predicados e objetos presentes na base de dados semântica:

```
SELECT ?subject ?predicate ?object
WHERE {
  ?subject ?predicate ?object
}
LIMIT 25
```

Figura 2.8: Exemplo da utilização de SPARQL

2.11 IBM Watson

Considera-se como Web Semântica a solução criada para que as máquinas consigam interpretar, manipular e processar as informações advindas de páginas da Internet, já que as informações depositadas no domínio da Web possuem o objetivo de serem destinadas para seres humanos e não computadores.

Outrora, caracteriza-se como computação cognitiva a arte de treinar as máquinas para que elas processem informações como seres humanos, logo é utilizado várias técnicas, tecnologias e ferramentas para que seja possível alcançar esse feitio, como por exemplo a utilização de aprendizado de máquina, redes neurais artificiais, inteligência artificial, probabilidade e estatística etc.

Levando em consideração o contexto de computação cognitiva, no ano de 2007 a IBM propôs o desenvolvimento de um software que possibilitasse a competição duma máquina com os campeões do jogo de perguntas e respostas denominado Jeopardy, (Gliozzo et al., 2011).

Desse modo, levaram-se quatro anos para a empresa apresentar um sistema que seja capaz de competir com os campeões de Jeopardy e que possua as peculiaridades necessárias para se sair vitorioso, como o processamento de uma exacerbada quantidade de informação, habilidades com processamento de linguagem natural, compreensão do que está sendo pedido e que determine com precisão e probabilidade a resposta correta, (Gliozzo et al., 2011).

O sistema que a IBM apresentou após quatro anos recebeu o nome de Watson e segundo Gliozzo et al. (2011), o software habituou-se com o jogo de

perguntas e respostas após a realização de uma quantidade extrema de partidas e em 2011 a máquina conseguiu realizar o feitio de vencer dos maiores campões do jogo, Ken Jennings e Brad Rutter.

O IBM Watson é baseado em uma arquitetura de software para análise de conteúdo de linguagem natural de fontes adversas de conhecimento denominada DeepQA, onde a mesma descobre, avalia e reúne todas as possíveis soluções para um determinado problema, sendo que suas fontes podem ser desde fontes estruturadas e não estruturadas até bases de conhecimento, como bancos de dados relacionais e não relacionais, Gliozzo et al. (2011). Sendo que, o software promove uma série de soluções provenientes do campo de aprendizado de máquina, processamento de linguagem natural, classificação de informação, Web Semântica e gerenciamento

O Watson está inserido como um serviço da plataforma de serviços da IBM denominada IBM Bluemix, sendo que está plataforma oferece diversas soluções para o gerenciamento de redes, desenvolvimento para aplicativos móveis, gerenciamento de apis e servidores, desenvolvimento de aplicações cognitivas etc.

Dessa forma, conforme o que foi explicitado anteriormente considera-se o IBM Watson um serviço da IBM que proporciona a outros sistemas a inserção de computação cognitiva em suas aplicações, logo neste trabalho de conclusão de curso será utilizado o mesmo aderido ao mecanismo de obtenção de informação líquida para proporcionar a arquitetura informacional proposta cognição.

Capítulo 3

TRABALHOS CORRELATOS

Este capítulo possui o objetivo de abordar artigos encontrados no estado da arte referente ao tema proposto neste trabalho, no qual baseia-se na elaboração de uma Arquitetura de Apoio a processos de Inovação baseada em estruturas informacionais semânticas e cognitivas.

3.1 Extração Resiliente de Dados RDF a partir de Fontes Dinâmicas em Linguagem de Marcação

Sabe-se que os dados presentes da internet, diversas vezes, não estão padronizados estruturalmente, logo informações referentes a atores de inovação em determinadas fontes informacionais possuem características de inserção de conteúdo adversas, portanto dificulta-se a obtenção destes através de mecanismos de extração de dados, pois os mesmos podem não entender a estrutura da notícia vigente, caso for modificado sua estrutura.

Desse modo, Mendonça (2013) propôs em seu trabalho uma forma de auto adaptação dos mecanismos de extração de dados, no qual caso o mecanismo falhe na obtenção da informação o mesmo reconstrói a configuração que guia o processo de extração, a partir do histórico de dados líquidos obtidos e num novo formato exportado pela fonte.

Dessa forma, além de propor uma solução para a problemática vigente deste trabalho a dissertação desenvolvida por este autor guia na elaboração de estratégias de obtenção de informações líquidas em fontes dinâmicas da WEB.

Além disso, atualmente grande parte das páginas da Web são feitas sintaticamente, ou seja, foram criadas para transmissão de informação para os seres humanos e não com a finalidade de manipulação e processamento desse conhecimento por parte dos computadores, portanto ao adicionarmos semântica aos agentes de captação de informação líquida de ambientes digitais informacionais contribui-se substancialmente para obtenção de uma informação mais inteligente e relevante sobre o conteúdo pesquisado.

Este trabalho foi selecionado como correlato, pois evidencia-se os problemas que existem para se extrair informações líquidas semânticas de web sites e o autor propõe uma solução baseada em reconstrução do guia do processo de extração. No entanto, neste trabalho de conclusão de curso é proposto uma solução adversa a de Mendonça (2013) que é a utilização de web crawlers, profundidade de extração e identificação de metadados e dados de interesse no espaço informacional digital.

Desse modo, conclui-se que esta pesquisa correlata agrega para este trabalho, pois ambos utilizam o mesmo meio que é a Web, retratam a extração de informação semântica e suas dificuldades para a obtenção, porém utilizam metodologias diferentes e técnicas diferentes.

3.2 Agente de Extração Informacional no Contexto de Big Data

Como foi explicitado no trabalho anterior, as páginas da Internet foram criadas de forma sintática por programadores e engenheiros, ou seja, foram construídas para o depósito de informação, no qual essa será lida por seres humanos. No entanto, com advento da Big Data são necessárias novas formas de estruturação da informação nas páginas da Web.

A partir desse contexto, o trabalho de Coneglian (2014) é o desenvolvimento de um agente de extração semântica no domínio da Web, no qual permite a localização, armazenamento, tratamento e recuperação de informações no contexto de Big Data. Inclusive, o autor elucida que seu trabalho serve como base

para a implementação de ambientes informacionais que auxiliam no processo de Recuperação de Informação.

Logo, nesse trabalho o autor além de propor um mecanismo de obtenção de informação líquida semântica no domínio da Web há o auxílio de uma ontologia que é responsável por aplicar a semântica ao processo de recuperação de informação. Não menos importante, como resposta do sistema de recuperação semântica há uma interface para a apresentação dos resultados obtidos aos usuários.

Coneglian (2014) para a implementação do robô de extração de dados informacionais utiliza a biblioteca denominada JSOUP feita em Java, outrora para a criação da ontologia emprega-se o software Protegé e a linguagem de elaboração de ontologias chamada de OWL. Portanto, após a confecção de sua ontologia o autor exporta a mesma para classes Java e assim consegue manipulá-la em conjunto com seu mecanismo de extração.

Coneglian (2014) em seu trabalho utiliza para a validação de sua plataforma a extração de informações específicas de artigos do site de IEEE Xplore (http://ieeexplore.ieee.org), no qual tais informações devem obedecer às regras estabelecidas por sua ontologia que foi criada conforme a metodologia de Noy e McGuiness (2001).

A metodologia de Noy e McGuiness (2001) possui a função de estabelecer o passo a passo necessário para a criação de uma ontologia, sendo que os sete passos são:

1. Determinar o domínio e o escopo da ontologia;
2. Reutilização de ontologias existentes;
3. Levantamento de termos importantes;
4. Definição de classes e suas hierarquias;
5. Definição das propriedades das classes;
6. Restrição de propriedades;
7. Criação de instâncias.

Portanto, considera-se este trabalho correlato, pois ambos utilizam o mesmo meio que é a Web, abordam e implementam ambientes de Big Data, além de compartilharem uma relação específica com a Web Semântica. No entanto, neste trabalho de conclusão de curso há algumas particularidades que divergem da pesquisa de Coneglian (2014), como a utilização de RDF e SPARQL, ou seja, outras

camadas da Web Semântica e a utilização de computação cognitiva como o IBM Watson.

Capítulo 4

PLATFORM TO SUPPORT INNOVATION

Neste capítulo será retratado qual a arquitetura informacional que engloba este projeto, o fluxograma informacional deste trabalho, principais funcionalidades e suas respectivas telas.

4.1 Arquitetura Informacional

Em uma pesquisa preliminar foi possível mapear os atores de inovação e seus produtos de informações que serviram como base para este trabalho. Nas figuras 4.1 e 4.2, destaca-se o espaço informacional identificado e as técnicas computacionais que serão utilizadas para a extração de informação com valor agregado para tomadas de decisão. Portanto, os módulos que serão o foco deste trabalho estão identificados a seguir:

- **Robôs Extratores**: mecanismo computacional responsável pela obtenção de informação líquida de ambientes informacionais digitais, sendo que trata-se de um Web Crawler capaz de capturar todos os metadados e dados de uma seção específica do domínio. Logo, como diferencial será a proposta de um robô capaz de preparar conteúdos para uma classificação semântica e posteriormente para uma análise cognitiva através do IBM Watson.
- **Ontologia:** as fontes informacionais exploradas neste trabalho são semiestruturadas, logo, a ontologia classificará semanticamente elas

utilizando a biblioteca Jena, ou seja, converterá informações de desestruturadas para estruturadas.

- **IBM Watson**: será utilizado neste projeto o IBM Watson para a realização de uma análise cognitiva das informações recuperadas.

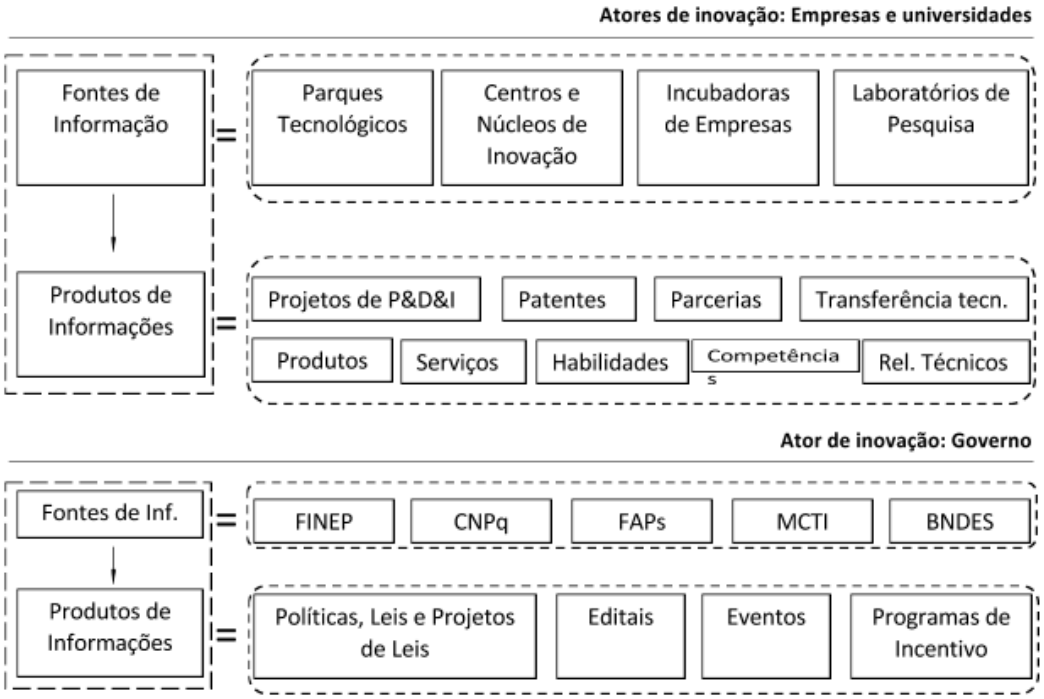


Figura 4.1: Arquitetura Informacional da Plataforma de Apoio à Inovação – Fonte: Adaptado de Pereira (2016, p.8)

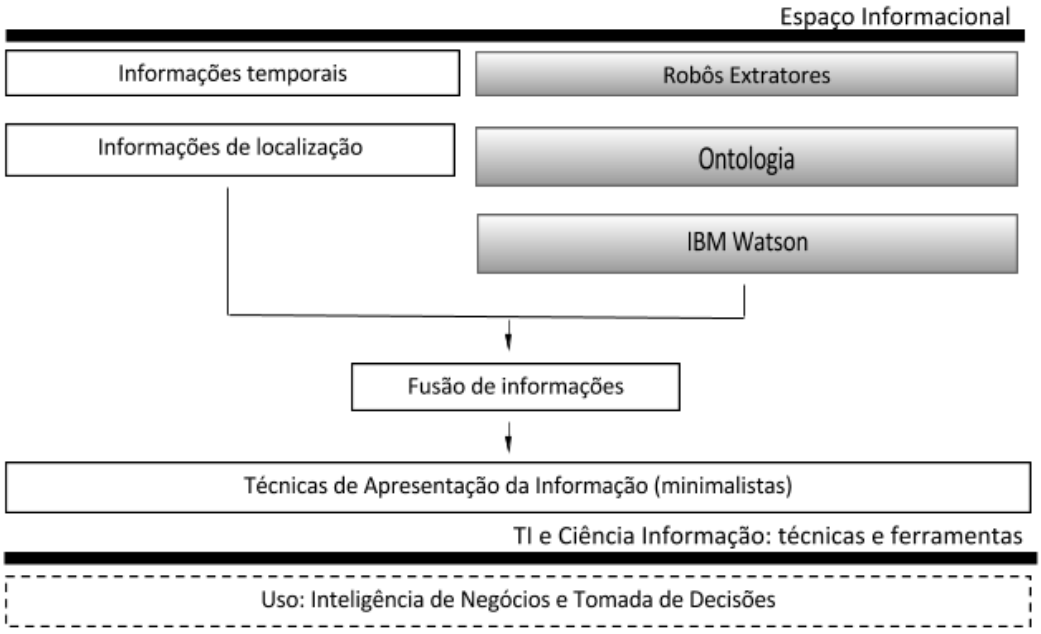


Figura 4.2: Continuação da Arquitetura Informacional da Plataforma de Apoio à Inovação – Fonte: Adaptado de Pereira (2016, p.8)

4.2 Platform to Support Innovation: Recovery of Semantic and Cognitive Information

Este trabalho teve como objetivo criar uma Plataforma de Apoio à Inovação com a capacidade de extrair, classificar e recuperar informações semânticas e cognitivas. Inclusive, foi nomeada de PLATSINN (*Platform to Support Innovation*), pois trata-se de uma plataforma que extrai metadados e dados de páginas Web e através de mecanismos de classificação e recuperação semântica e cognitiva possibilita ao usuário realizar tomadas de decisão em prol de melhorar a competitividade no mercado através da Inovação.

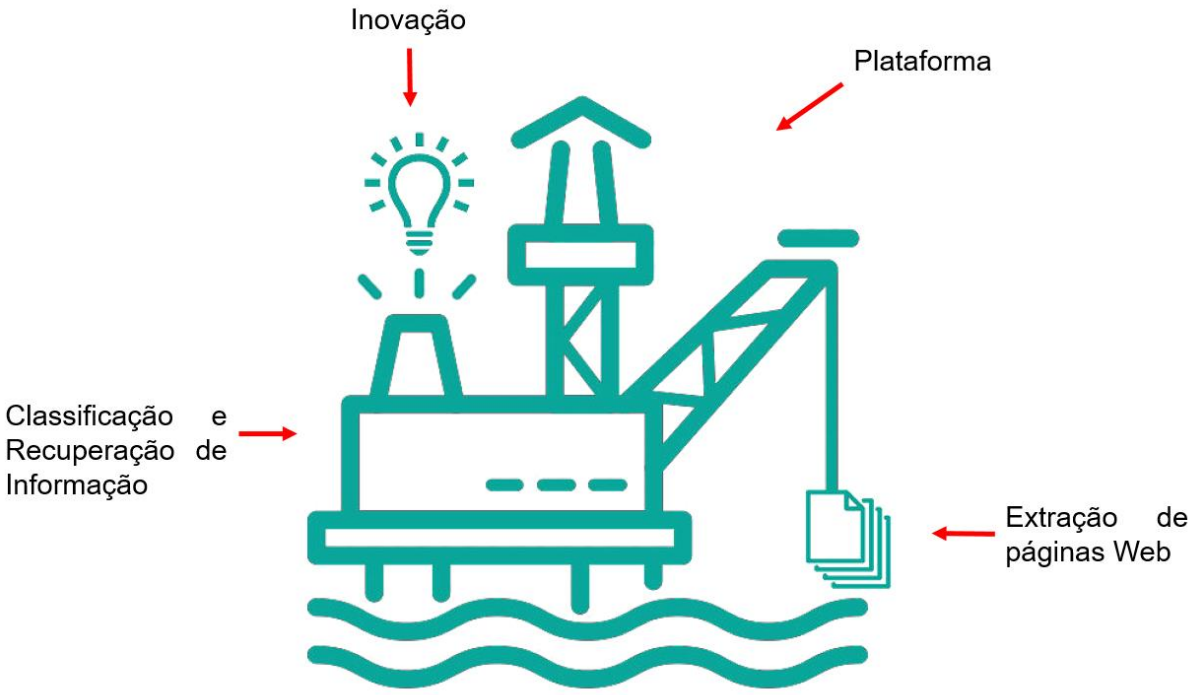


Figura 4.3: Logo do PLATSINN explicada

Além disso, em seguida há um fluxograma informacional de como o PLATSINN comunica com os seus módulos:

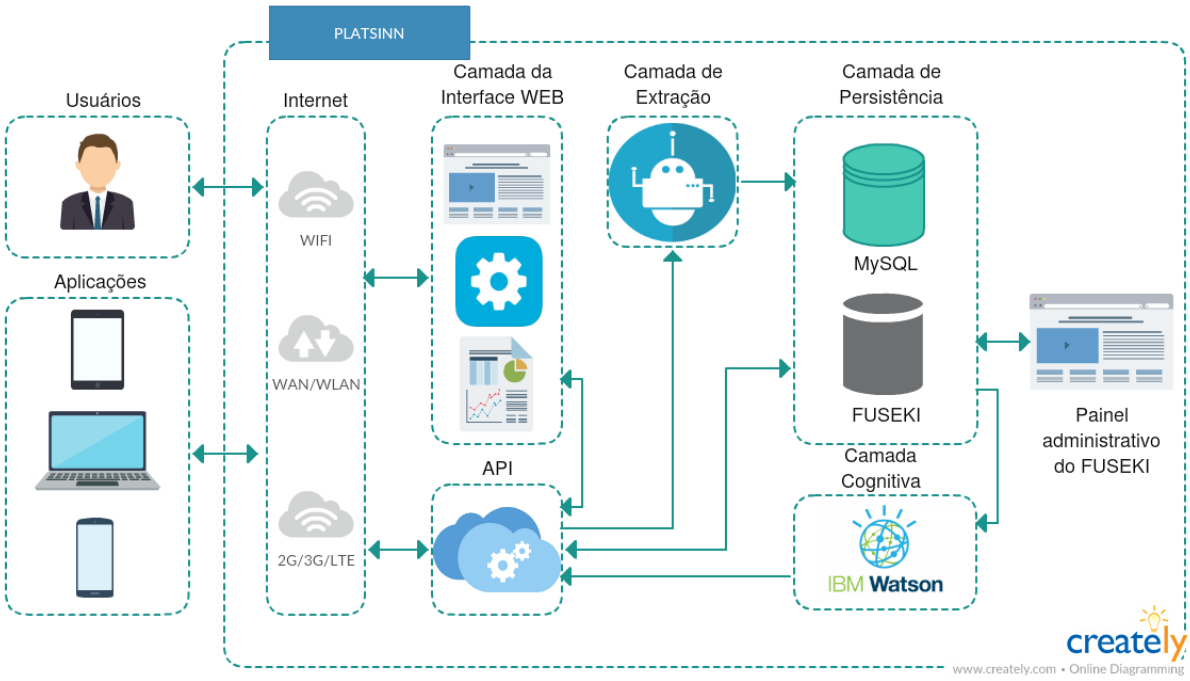


Figura 4.4: Fluxograma Informacional do PLATSINN

O PLATSSIN como foi explicitado anteriormente tem a capacidade de obter, especificar e filtrar informações semânticas e cognitivas somente de notícias, editais ou eventos. Portanto, conforme o fluxograma informacional exibido na figura 4.4, caso um cliente deseje realizar uma extração na Plataforma de apoio à Inovação basta o usuário ou uma aplicação requisitar à “Camada da Interface Web” ou diretamente a “API” e posteriormente os dados serão percorridos através da “Camada de Extração” e “Camada de Persistência”. Desse modo, o resultado chegará ao cliente ou ao serviço externo através do módulo da “API” ou “Camada da Interface Web”. Sendo que, o formato de dado que percorre entre os módulos é o JSON.

Inclusive, quando o usuário solicita o processo de extração automaticamente os metadados e dados obtidos são transformados em informação, classificados semanticamente na ontologia InovaOnto com o apoio da biblioteca Jena e persistidos no servidor FUSEKI através de serviços de conexão HTTP.

Ademais, segundo o fluxograma informacional exposto na figura 4.4, se o cliente desejar recuperar informações basta requisitar à “Camada da Interface Web” ou diretamente a “API” e os dados da solicitação percorrerão até a “Camada de Persistência” e serão processados pelos mecanismos de filtro informacional. Dessa

forma, a resposta chegará ao cliente ou ao serviço externo através do módulo da “API” ou “Camada da Interface Web” em formato JSON.

Além disso, para a realização de uma análise cognitiva em alguma notícia, edital ou evento, primeiramente, o cliente precisa requerer a informação à “Camada da Interface Web” ou diretamente a “API” e os dados da solicitação percorrerão até a “Camada de Persistência” e serão processados pelos mecanismos de filtro informacional. Logo, o resultado obtido da “Camada de Persistência” será processado pela “Camada Cognitiva” e retornará uma resposta no formato JSON com categorias, entidades, emoções, sentimentos, conceitos e palavras-chave do texto analisado.

O PLATSINN em sua interface Web é dividido em três categorias (“Configurações”, “Base de Informações” e “API”), onde em cada uma há operações que podem ser realizadas, portanto a seguir encontra-se uma imagem da tela principal da plataforma com cada categoria e funcionalidade discriminada:

Figura 4.5: Tela principal da interface Web do PLATSINN

Conforme pode ser observado na figura 4.5 as categorias e funcionalidades da Plataforma de apoio à Inovação são:

1. **HOME:** A página principal da interface Web do PLATSINN, onde nela há estatísticas sobre a plataforma, como: “Relação quantitava entre Instâncias Estáticas e Dinâmicas”; “Relação quantitava entre Tipos Estáticos e Dinâmicos”; “Relação quantitava de extrações por domínio”; “Relação quantitava de Links Informacionais persistidos por domínio”; “Relação quantitava de informações advindas de entidades públicas e privadas” e “Relação quantitava de informações advindas de eventos, notícias e editais”. Inclusive, ao clicar sobre o ícone da categoria o usuário é redirecionado para “home.php”, ou seja, a própria página.
2. **CONFIGURAÇÕES:** Nesta categoria é disponibilizado as seguintes funções:
    a. **Cadastrar Extração:** Nesta função ao clicar o cliente será redirecionado para “cad_extracao.php” que possibilita ao utilizador da plataforma cadastrar uma extração informando parâmetros, como: “link”; “profundidade” e “tipo dinâmico”.

        Portanto, a seguir segue uma imagem da tela de cadastro de extração confeccionada para o PLATSINN:

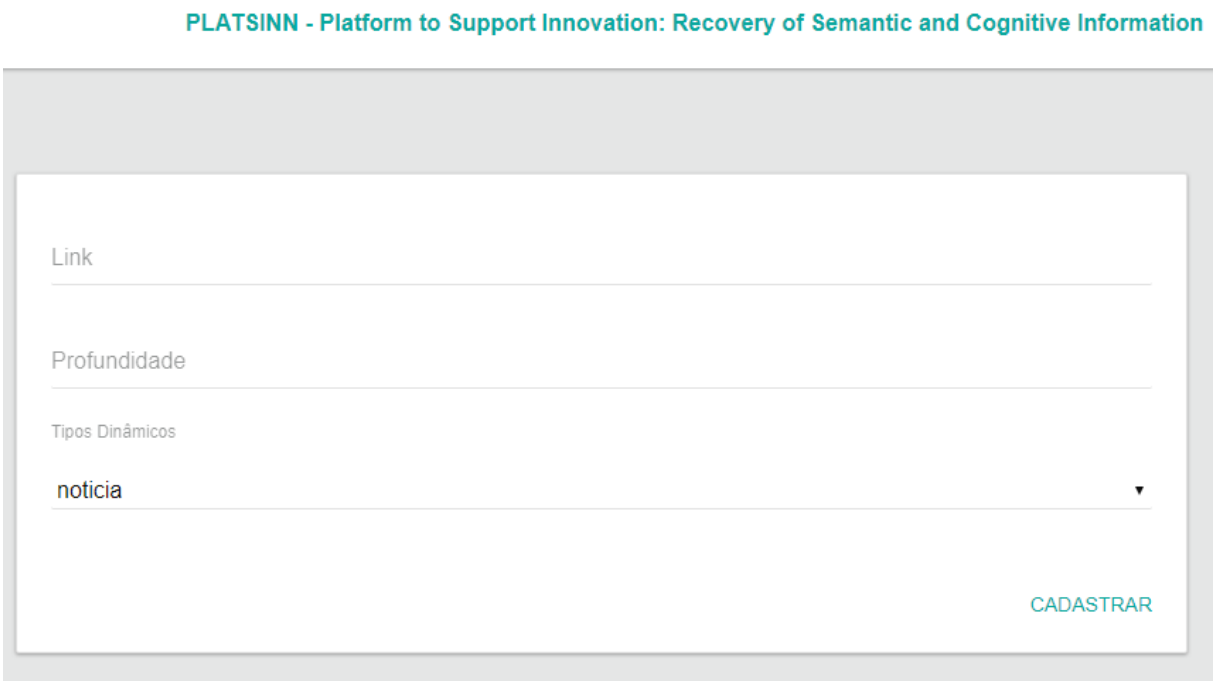


Figura 4.6: Tela de cadastro de extração

b. **Deletar Extração:** Ao acessar esta funcionalidade o usuário é migrado para a página “del_extracao.php” que proporciona a exclusão das informações obtidas com o robô de busca .

Logo, a seguir existe uma imagem da tela de exclusão de extração:

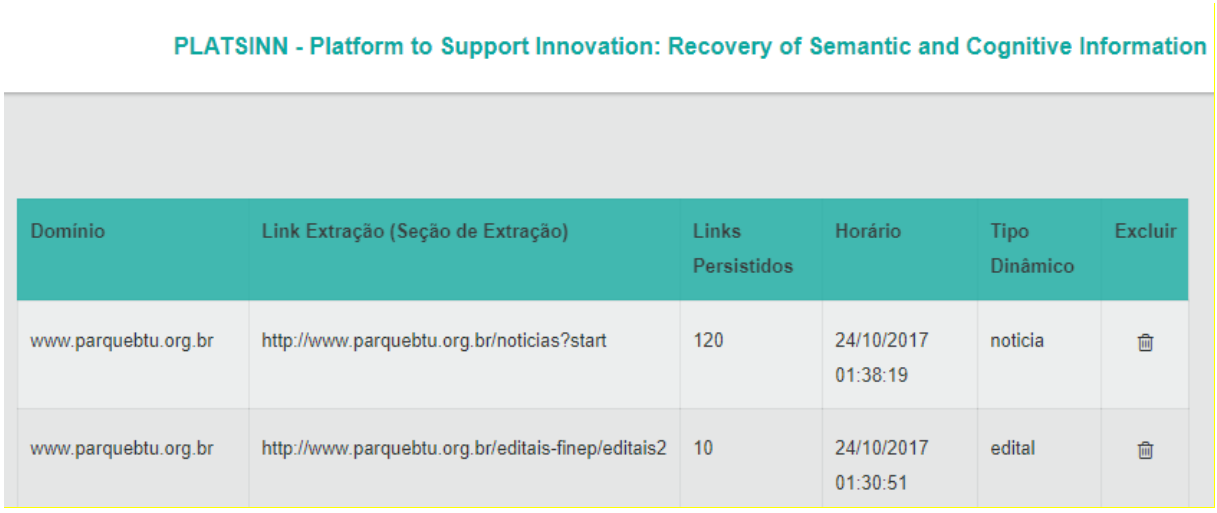


Figura 4.7: Tela de exclusão do PLATSINN

Desse modo, na figura 4.7 é vísivel que o PLATSINN traz algumas informações sobre extrações para que o usuário tome uma decisão se realmente necessita excluí-la, como respecivamente: domínio da extração; seção que foi extraída; quantidade de links persistidos; horário que foi realizada a obtenção dos metadados e dados; qual é o tipo de informação que foi obtida; ação de exclusão.

c. **Cadastrar Instancia Estática:** Nesta subcategoria ao clicar o usuário será redirecionado para “cad_instestatica.php” que proporciona ao utilizador da plataforma cadastrar uma instância estática informando parâmetros, como: “nome”; “site”, “tipo estático” e “tipo entidade”;

Logo, em seguida há uma imagem da tela de cadastro de instância estática confeccionada para o PLATSINN:

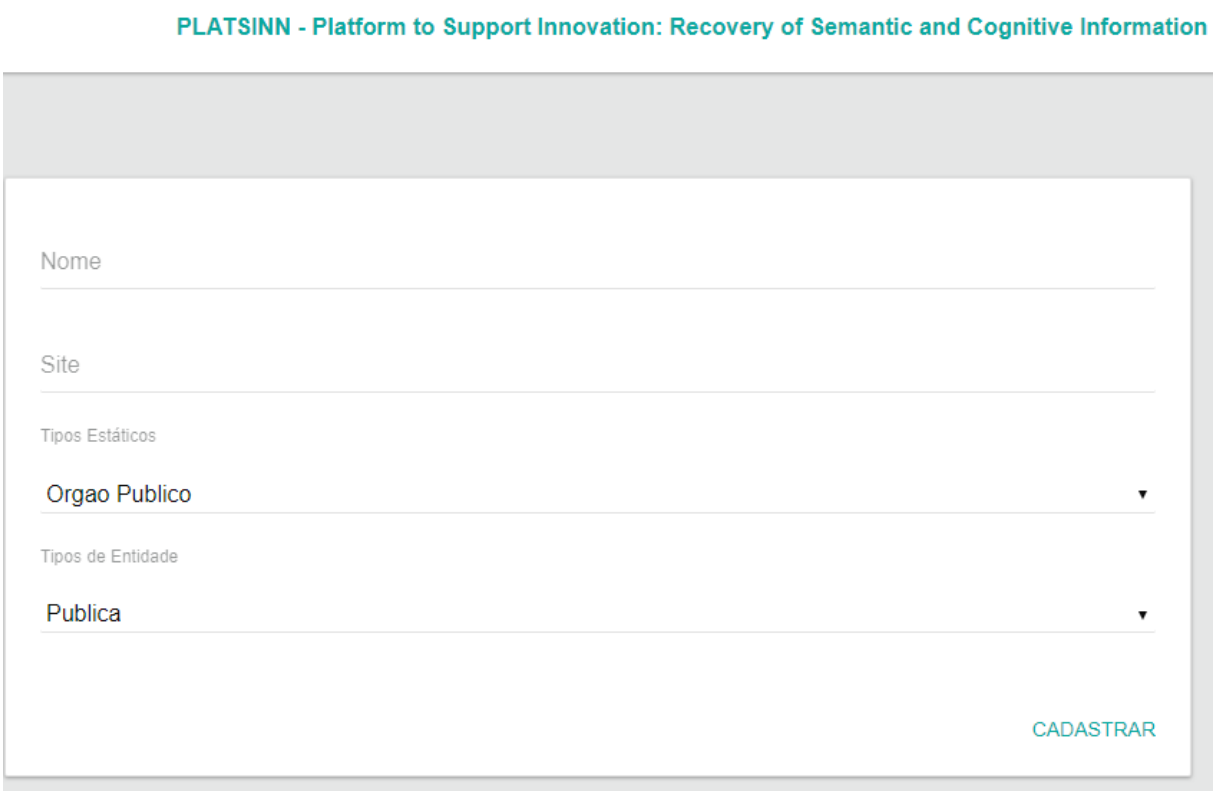


Figura 4.8: Tela de cadastro de instância estática do PLATSINN

d. **Deletar Instancia Estática:** Ao acessar esta subcategoria o cliente é migrado para a página “del_instestatica.php” que proporciona a exclusão de uma instância dinâmica.

Logo, a seguir existe uma imagem da tela de exclusão de instância estática:

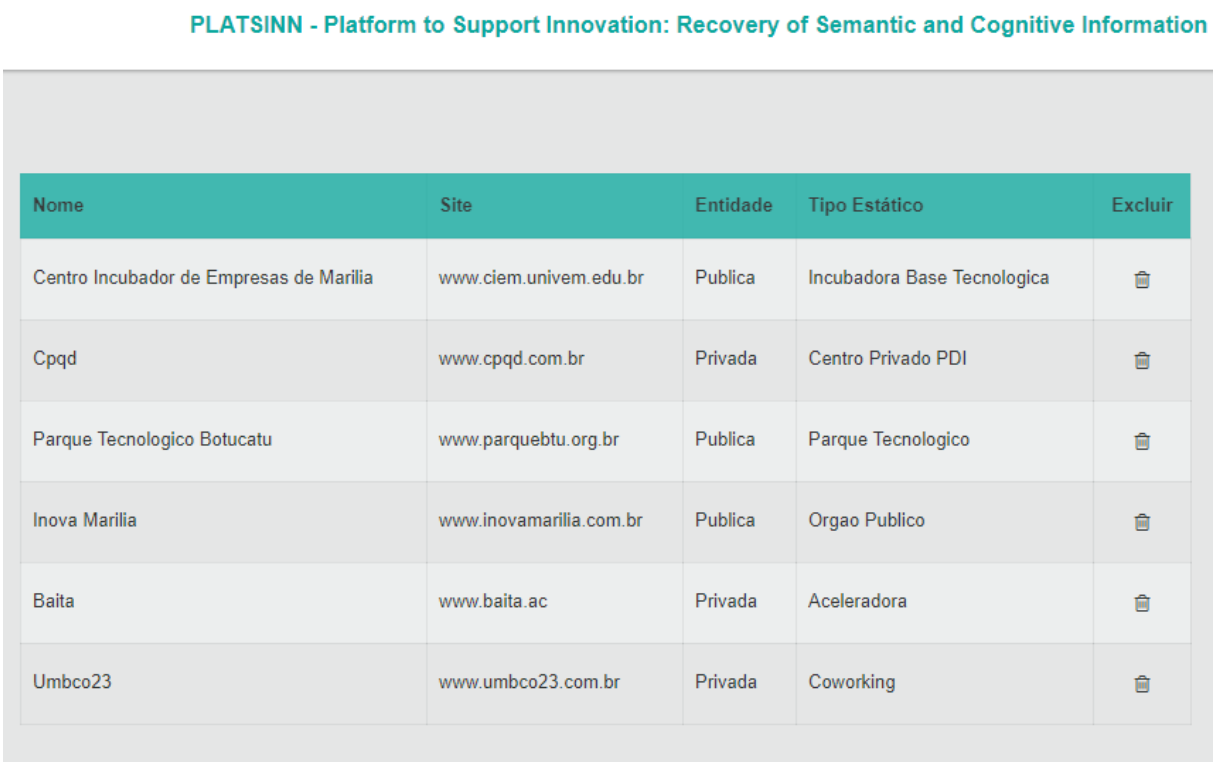


Figura 4.9: Tela de exclusão de instância estática do PLATSINN

Desse modo, na 4.9 é possível observar que a plataforma traz algumas informações sobre as instâncias estáticas, como respecivamente: nome; site; entidade pública ou privada; qual é o seu tipo (aceleradora, coworking, Parque Tecnologico)

3. BASE DE INFORMAÇÕES:

a. **Ontologia:** Ao clicar nesta funcionalidade o usuário é redirecionado para a página “inovaonto.php”, onde é exibido informações sobre a ontologia utilizada na plataforma.

b. **Fuseki:** Ao acessar esta subcategoria o cliente será redirecionado para “fuseki.php”, no qual exibirá informações sobre o servidor FUSEKI.

c. **Filtro Informacional:** Nesta subcategoria o cliente ao acessá-la é redirecionado para uma página chamada “filtro_fuseki.php” que possibilita filtrar notícias, editais ou eventos, logo a imagem desta tela pode ser observada a seguir:

Figura 4.10: Tela de filtro informacional do PLATSINN

Na tela apresentada na figura 4.10 os parâmetros necessários para a filtragem de notícias, eventos ou editais são: “domínio”, “tipo dinâmico” e “quantidade de informações exibidas”.

Logo, abaixo é possível observar a imagem da tela resultante do filtro informacional denominada “filtro_fuseki_sparql.php” criada para o PLATSINN:

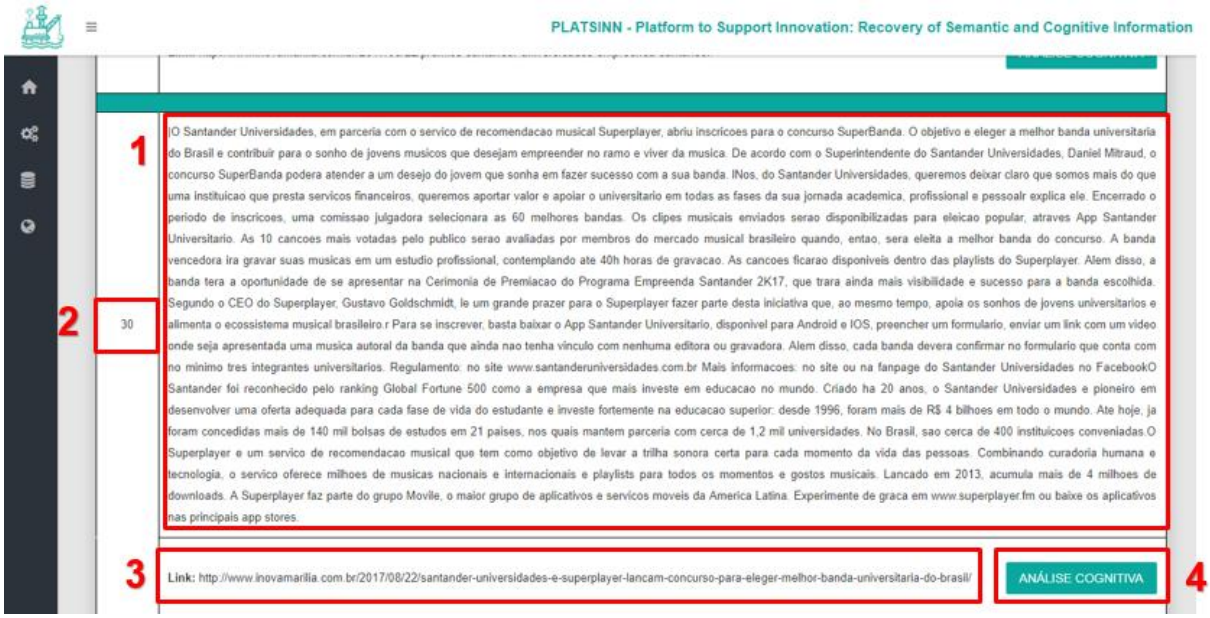


Figura 4.11: Tela resultante do filtro informacional do PLATSINN

Inclusive, conforme a figura 4.11 as informações resultantes apresentam características, respectivamente:

1. Contéudo da notícia, edital ou evento;
2. Índice da informação;
3. Endereço eletrônico;
4. Botão de análise cognitiva do conteúdo apresentado.

Além disso, a seguir observa-se a imagem da tela de análise cognitiva chamada de “filtro_watson.php”, no qual é acionada quando o usuário deseja obter informações cognitivas sobre o conteúdo apresentado no filtro informacional:

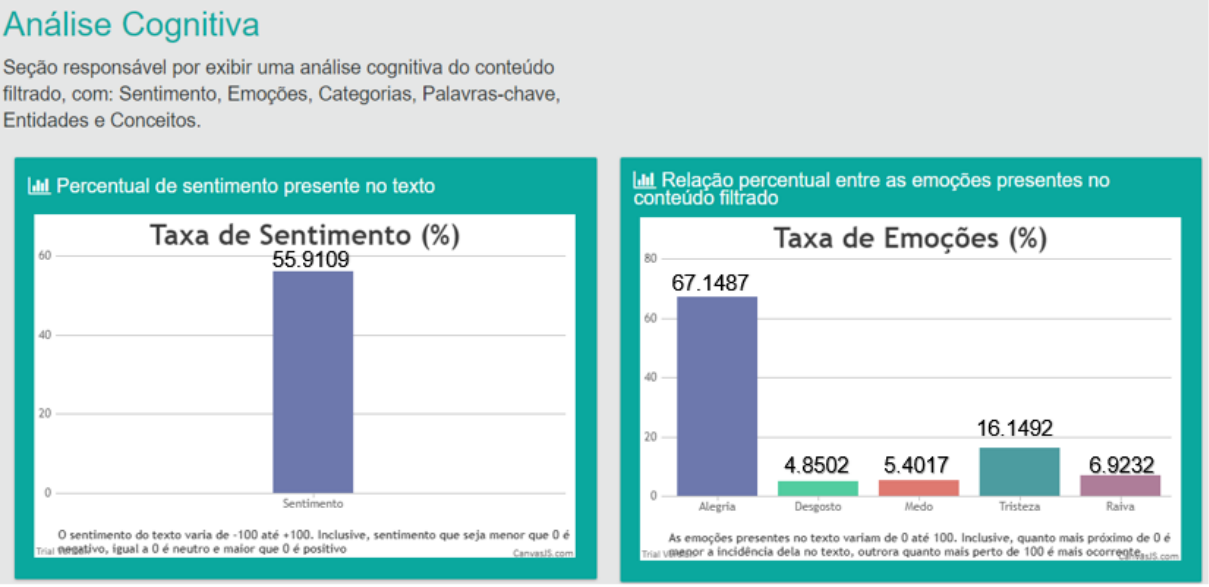


Figura 4.12: Tela de análise cognitiva do PLATSINN

Na página da figura 4.12 apresenta-se: “Percentual de sentimento presente no texto” em formato de gráfico de colunas ; “Relação percentual entre as emoções presentes no conteúdo filtrado” com uma representação gráfica em colunas; “Categorias” através de uma lista; “Conceitos” por meio de uma listagem; “Entidades” através de uma lista e “Palavras-chave” através de uma listagem.

d. **Filtro Informacional Avançado:** Ao acessar esta subcategoria o usuário é redirecionado para a página de consulta da base de dados “filtro_avancado_fuseki.php”, no qual possibilitará o cliente realizar consultas SPARQL avançadas.

4. **API**

a. **PLATSINN:** Nesta subcategoria o usuário ao clicar é redirecionado para “platsinn_webservice.php” que exibi informações das funcionalidades presentes na API confecionada.

Capítulo 5

DESENVOLVIMENTO

Este capítulo possui o objetivo de descrever como foi desenvolvida a Plataforma de Apoio à Inovação (PLATSINN), inclusive nele será retratado o processo de implementação de cada parte: extração de metadados e dados, transformação dos dados coletados em informação, alteração da ontologia InovaOnto, classificação e persistência semântica das informações, classificação cognitiva, recuperação da informação, API, etc.

5.1 Espaço Informacional

Primeiramente, este Trabalho de Conclusão de Curso utilizou como espaço informacional sites que retém um acervo considerável de informações referentes à inovação, pesquisa científica, parques tecnológicos, centros de inovação, etc. Logo, os sites utilizados e suas respectivas seções extraídas podem ser observados na lista a seguir:

- Inova Marília (www.inovamarilia.com.br):
    - http://www.inovamarilia.com.br/category/noticias/
    - http://www.inovamarilia.com.br/category/editais/
    - http://www.inovamarilia.com.br/category/eventos/
- CIEM – Centro Incubador de Empresas de Marília(www.ciem.univem.edu.br):

    - http://ciem.univem.edu.br/noticias/
    - http://ciem.univem.edu.br/eventos/
- Parque Tecnológico de Botucatu (www.parquebtu.org.br):
    - http://www.parquebtu.org.br/noticias?start
    - http://www.parquebtu.org.br/editais-finep/editais2
- Baita Aceleradora (www.baita.ac):
    - http://www.baita.ac/category/news/
- Umb.com23 – Escritório Compartilhado (www.umbco23.com.br):
    - http://umbco23.com.br/blog/
- FAPESP (http://www.fapesp.br/):
    - http://www.fapesp.br/secao/noticias

5.2 Extração de metadados e dados

A plataforma de Apoio à Inovação possui como pilar de sua aquisição de informação um robô de busca capaz de obter metadados e dados de ambientes informacionais digitais. Desse modo, o mecanismo computacional faz uma busca no HTML das páginas de uma seção de um domínio por *tags* predefinidas, ou seja, caso o agente de extração encontre um link que esteja dentro do espaço informacional desejado e numa determinada profundidade (paginação) as informações líquidas adquiridas são transformadas em cadeias de caracteres.

Dessa forma, a seguir é possível observar o processo do mecanismo de obtenção de informação líquida:

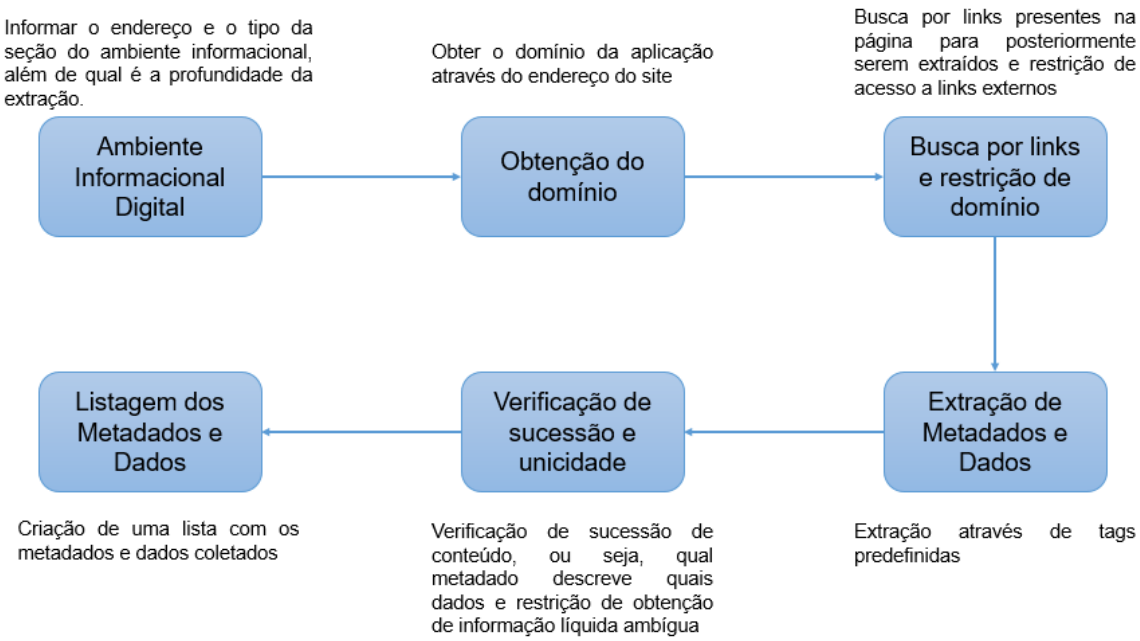


Figura 5.1: Fluxograma da Extração de informações

Como mostrado na Figura 5.1 o agente de extração é dividido em 6 etapas denominadas: ambiente informacional digital, obtenção do domínio, busca por links e restrição de domínio, extração de metadados e dados, verificação de sucessão e unicidade e listagem dos metadados e dados.

- **Ambiente informacional digital:** Primeiramente, nesta etapa caracteriza-se pela definição de qual seção do espaço informacional será realizado a extração, qual a profundidade que o mecanismo deverá percorrer e se o conteúdo obtido classifica-se em: notícia, edital ou evento. Desse modo, a seguir encontra-se uma imagem ilustrativa de como reconhecer tais parâmetros numa página.

Figura 5.2: Exemplo de reconhecimento dos parâmetros de extração – Fonte: Adaptado do site do Inova Marília (http://www.inovamarilia.com.br/category/noticias/)

Desse modo, conforme a figura 5.2 observa-se que todas as páginas que contém o endereço eletrônico da seção (http://www.inovamarilia.com.br/category/noticias/) terão a profundidade zero, pois o robô de busca não necessitou acessar outros endereços, já que seus links podem ser obtidos na área de paginação, por exemplo:

- http://www.inovamarilia.com.br/category/noticias/page/2/
- http://www.inovamarilia.com.br/category/noticias/page/3/
- http://www.inovamarilia.com.br/category/noticias/page/4/

Além disso, os endereços eletrônicos que são resultantes do acesso e extração de outras páginas têm a profundidade acrescida para cada acesso que o mecanismo de extração necessitou realizar.

Por exemplo, o agente de extração obteve informações do seguinte endereço: “http://www.inovamarilia.com.br/2017/07/14/parque-tecnologico-botucatu-pretende-desenvolver-acoes-em-parceria-com-centro-de-inovacao-de-marilia/”, no entanto, este link foi conquistado através da aquisição de metadados e dados presentes nesta seção deste site: “http://www.inovamarilia.com.br/category/noticias/page/3/”, onde ao ser extraída possuía como profundidade 0. Logo, como existiu somente uma

conexão do robô de extração entre a página da categoria e a notícia, portanto a profundidade dela é 1.

Dessa forma, conclui-se que para efetuar uma extração com o mecanismo computacional criado neste trabalho de conclusão de curso é imprescindível atentar-se em qual profundidade encontra-se o conteúdo solicitado. Sendo que, geralmente notícias, editais e eventos situam-se em profundidade de nível 1 nos sites do escopo deste trabalho, entre outros.

- **Obtenção de domínio:** O agente pegará o link informado e extrairá o domínio do espaço informacional, por exemplo, para o link da seção de notícias do site Inova Marília (http://www.inovamarilia.com.br/category/noticias/) o domínio é (www.inovamarilia.com.br).
- **Busca por links e restrição de domínio:** a aplicação desenvolvida trata-se de web crawler, logo este mecanismo possui a característica de ao encontrar um hiperlink guardá-lo em uma lista de conexões futuras e repetir esta operação diversas vezes, no entanto, isso pode acarretar um loop infinito caso não seja tratado quais links ele pode acessar. Por isso, é imprescindível que a extração seja restringida ao domínio do ambiente informacional e a uma verificação se o endereço já foi acessado a partir de uma lista de conexões efetuadas.
- **Extração de metadados e dados:** o robô de busca tende a procurar por hiperlinks que respeitem sua política de acesso a endereços Web para guardar em uma lista e acessá-los futuramente, onde só é possível conectar aqueles que estiverem no mesmo domínio, na profundidade definida e não serem inibidos pela expressão regular de restrição (“.*\\.(png|jpg|gif|bmp|pdf|ppt|pptx|jpeg|xml|csv)”), nesta expressão é aceito todos os links que não tiverem o término descriminado entre barras.

As tags do tipo head (h1, h2, h3, h4, h5, h6) segundo o W3C (World Wide Web Consortium) são utilizadas em cabeçalhos de páginas Web, dessa forma, levando em consideração que metadados são dados que descrevem dados e no contexto desta pesquisa essas tags servem para descrever o conteúdo que será explicitado posteriormente a sua aparição, logo conclui-se que essas tags são metadados desestrurados..

Inclusive, nesta pesquisa adota-se como estrutura HTML específica provedora de conteúdo caracterizado como dado as tags “p”, “pre”, “span”, “i”,

“strong”, “a”, em que, tais estruturas representam respectivamente: parágrafo, texto pré-formatado, elemento estilizado, itálico, negrito, endereço eletrônico. Na figura a seguir demonstra-se numa página tal relação:

Figura 5.3: Exemplo de página com metadados e dados – Fonte: Adaptado do site do Inova Marília (http://www.inovamarilia.com.br/2017/05/04/univem-realiza-palestra-sobre-startups-e-aceleradoras/)

Dessa forma, foi criado uma classe Java chamada “Dado” que encapsula todas as informações necessárias para prover metadados e dados ao usuário do mecanismo de extração e seus atributos são:

- O tipo da informação extraída (metadado ou dado);
- A tag que o conteúdo se encontra;
- O conteúdo;
- O endereço eletrônico do conteúdo extraído do espaço informacional;
- Em qual profundidade o metadado se encontra;
- O CSS do contéudo.

As informações coletadas são convertidas para o formato String e instanciadas em um objeto Java da classe explicitada anteriormente. Posteriormente, são acrescentados esses dados numa lista para que seja possível a realização da verificação de sucessão de cada metadado e dado.

- **Verificação de sucessão e unicidade:** Ao obter as informações líquidas e adicioná-las em uma lista de dados é preciso criar um mecanismo que modele

e relacione os metadados e os dados. Por exemplo: o metadado da 5.3 intitulado “Univem realiza palestra sobre startup e aceleradoras” descreve uma notícia, onde ela contém dados, como: “O Univem por meio dos cursos [...]”, etc. Dessa forma, criou-se uma classe chamada “MDado” que possui como atributos dois objetos da classe “Dado”, ou seja, um relacionamento um para um de metadados e dados.

Logo, a seguir observa-se o diagrama UML das classes “Dado” e “Mdado”:

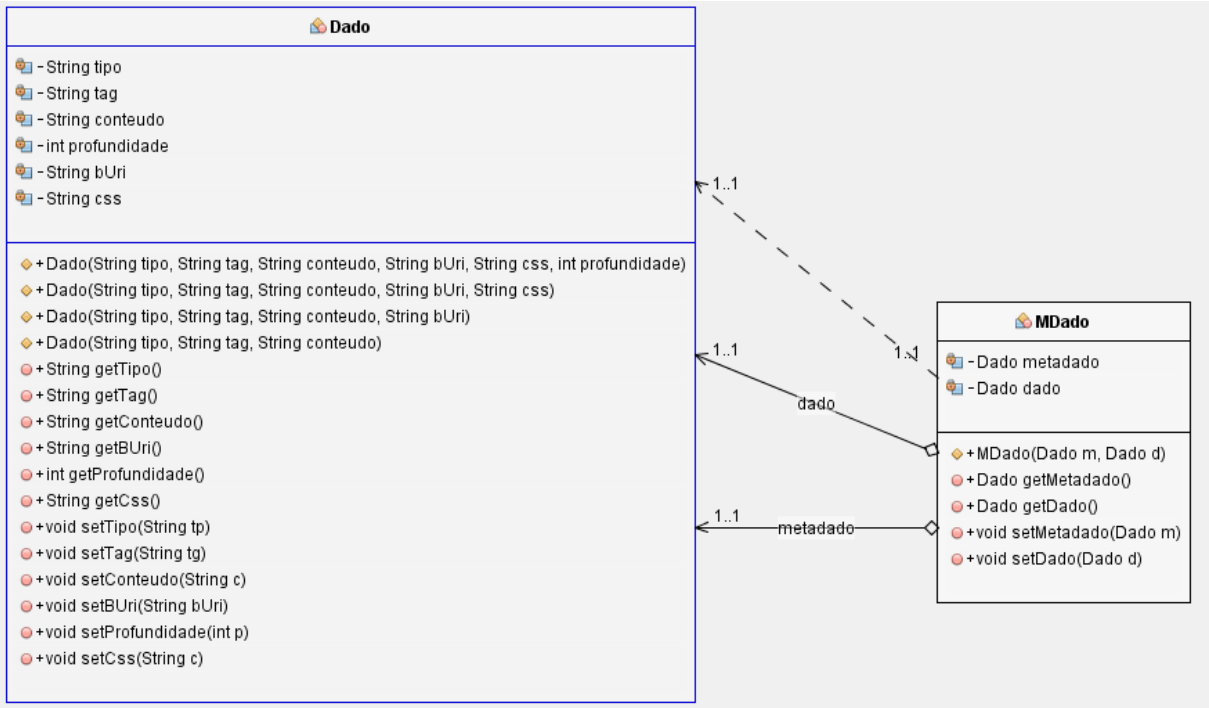


Figura 5.4: Diagrama UML das classes responsáveis pelo molde da informação extraída

No entanto, como é feito a parte do relacionamento? Primeiramente, foi elaborado um código em Java capaz de identificar quais metadados e dados se relacionam através de repetidos processos de verificação de sucessão.

A verificação de sucessão é feita por três estruturas de controle capazes de comparar o elemento atual com o posterior. Ademais, nela existe um acréscimo de uma averiguação de unicidade, no qual baseia-se na busca pela inexistência de outros objetos com conteúdos idênticos na lista de “MDado”, portanto a seguir segue uma explanação:

- Se o tipo do elemento é um “metadado” e o próximo é um “metadado”:

- É criado uma instância de “MDado”, onde os atributos “dado” e “metadado” são do próprio elemento. Por exemplo: Na imagem a seguir observa-se que há dois títulos descrevendo a notícia, onde ambos possuem o mesmo conteúdo. No entanto, como definir qual é a relação com as demais informações presentes na página? A solução é constatada da seguinte maneira, o primeiro título denominado “Innova Space Coworking [..]” tem como tag um h3 e o posterior também, logo ao instanciar o objeto “MDado” com as informações do inical o atributo “metadado” será a tag dele e o “dado” corresponderá ao próprio objeto.

Figura 5.5: Exemplo sucessão de metadados – Fonte: Adaptado do site do Inova Marília (http://www.inovamarilia.com.br/2017/10/17/innova-sapce-coworking-recebe-alunos-do-ensino-medio/)

- Caso o tipo de dado seja um “metadado” e o seguinte seja classificado como “dado”, logo é instanciado vários objetos “MDado” com o dado classificado como “metadado”:
  - Será instanciado um objeto da classe “MDado”, onde o atributo “metadado” corresponderá ao próprio elemento e o “dado” serão os elementos posteriores. Sendo que, as construções sucessivas de objetos com esta característica é inibida quando encontrar um “metadado” ou exceder o tamanho da lista de elementos.

- No caso de ambos sejam dados:
  - É criado uma instância de “MDado”, onde os atributos “dado” e “metadado” são respectivamente: os elementos posteriores até ser encontrado um dado do tipo “metadado” e o próprio elemento.

- **Listagem dos Metadados e Dados:** As informações confecionadas são inseridas numa lista de “MDado” que possui a função de conter todos os metadados com seus respectivos dados para posteriormente ser feito uma classificação semântica.

5.3 Transformação em conteúdo

O PLATSINN após realizar o processo de extração de metadados e dados de páginas Web precisa transformar isso em conteúdo valorado, pois, logo após este processo é realizado a classificação e persistência semântica deste conteúdo com o auxílio da biblioteca JENA e do servidor FUSEKI.

Desse modo em seguida, é apresentado um fluxograma da como é feito este processo:

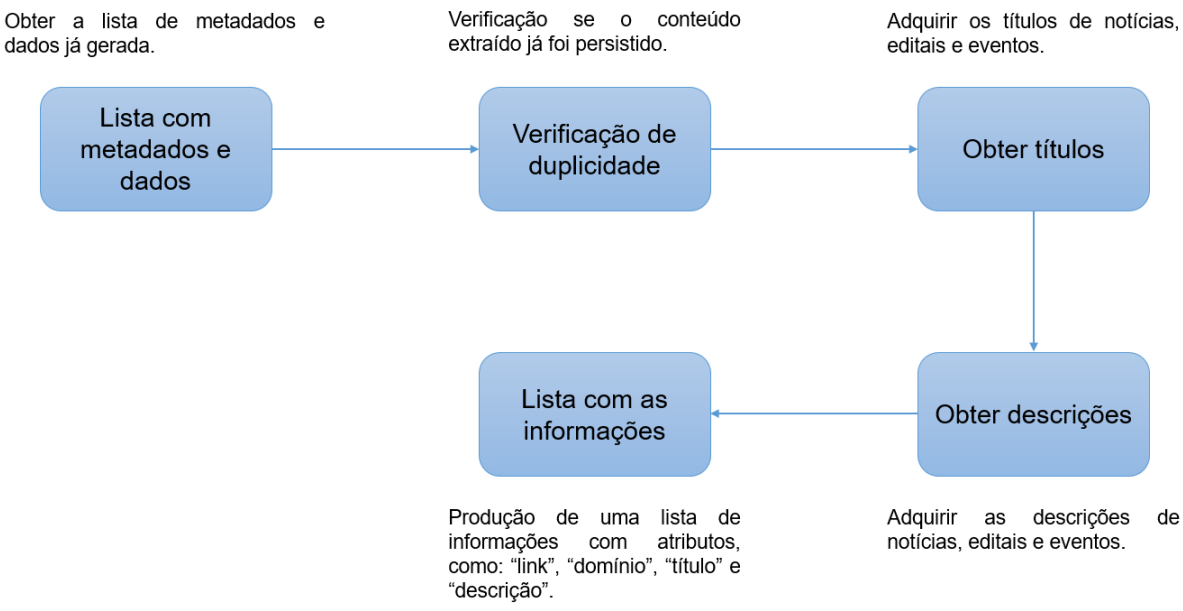


Figura 5.6: Fluxograma da transformação de metadados e dados em informação

Como mostrado na 5.6 o este processo é dividido em 5 partes denominadas: “Lista com metadados e dados”; “Verificação de duplicidade”; “Obter títulos”; “Obter descrições”; “Lista com as informações”.

- **Lista com metadados e dados:** Primeiramente, nesta etapa o mecanismo de transformação em informação receberá a lista de metadados e dados obtida com o robô de busca;
- **Verificação de duplicidade:** Ao adquirir esta lista o mecanismo transformará o conteúdo presente em JSON através da biblioteca GSON e conveterá o resultado para Hash do tipo “SHA-256” com o auxílio da ferramenta MessageDigest. Dessa forma, o PLATSINN consultará se existe alguma extração realizada que possua o “link” e “hash” idênticos, pois se houver o processo de classificação e persistência semântica não será realizado, já que a plataforma tem o intuito de evitar a ambiguidade informacional.
- **Obter títulos:** A obtenção de títulos de notícias, editais ou eventos é realizado da através de um vasculhamento na lista de metadados procurando aqueles que possuem tags de descrição de título (“h1”, “h2”, “h3”, “h4”, “h5”, “h6”). Desse modo, se objeto referente a página extraída possuir somente um metadado de descrição de título, logo, ele se torna a titulação da página. No entanto, se houver adversos metadados candidados para a titulação o mecanismo desta etapa realiza uma busca por sucessão de metadados através de uma ordem de prioridade (h1 > h2 > h3 > h4 > h5 > h6) ou por meio de uma busca pelo metadado descritor de dados mais frequente na página.
- **Obter descrições:** Esta etapa caracteriza-se por armazernar em uma veriável todos os dados da lista de metadados e dados que contém conteúdo de uma página.
- **Lista com as informações:** Após a realização desta conquista de títulos e descrições de ambientes informações digitais basta armazená-los numa lista para posteriormente classificá-los semânticamente na ontologia Inova Onto.

5.4 Alteração da Ontologia InovaOnto

Primeiramente, como o escopo desta Plataforma de Apoio à Inovação restringe-se a atores de inovação (empresas, universidades e governo) que possuem alguma relação com Parques Tecnológicos e Centros de Incubadores de Empresas credenciados no Sistema Paulista de Ambiente de Inovação (SPAI), logo necessitou-se de um modelo informacional com o intuito de tratar a informação adquirida pelo mecanismo de obtenção de informação líquida. Portanto, utilizou-se uma ontologia conceitual criada por Fusco & Mucheroni & Coneglian (2017), onde possui o objetivo de abstrair o cenário de inovação do SPAI.

Fusco & Mucheroni & Coneglian (2017) criou um modelo ontológico e um diagrama UML capaz de ajudar a compreender o ecossistema de inovação do SPAI, onde ressalta quais são os principais atores, ambientes de inovação, entidades de fomento, entidades de apoio à inovação, empresas, além dos respectivos relacionamentos com editais, eventos, legislação, projetos, pesquisadores, fundos de amparo, aceleradoras, órgãos públicos, centros privados de pesquisa e desenvolvimento etc.

No entanto, Fusco & Mucheroni & Coneglian (2017) em seu artigo além de ter produzido um modelo conceitual não colocou um relacionamento entre as informações advindas de notícias com o ecossistema de inovação do SPAI, consequentemente em seu UML inexistia propriedades que garantissem a inserção de informações midiáticas em sua ontologia e não aderiu propriedades a todas as classes no seu modelo UML. Portanto, exigiu-se uma modificação da UML e do OWL confeccionado pelo autor.

As modificações no campo ontológico são relacionadas a conceitos de instâncias estáticas e dinâmicas. No qual, as estáticas são aquelas advindas de entidades em que as modificações informacionais são bem raras, portanto, não há a necessidade de um robô de busca para a aquisição desses conhecimentos, pois pode-se realizar esse processo manualmente com o apoio de um programa ou ferramenta de manipulação de OWL. Outrora, instâncias dinâmicas são aquelas provenientes de entidades que as modificações informacionais são constantes, como: notícias, editais, eventos, legislação, projetos, pesquisadores etc. Dessa forma, necessita-se de um mecanismo de obtenção de informação líquida que

promova o acréscimo de conhecimento de forma automática sem o intermédio do humano da hora da persistência numa base de dados semântica.

As modificações realizadas no trabalho de Fusco & Mucheroni & Coneglian (2017) que capacitou a Plataforma de Apoio à Inovação recuperar informações semânticas foram:

- **Definição das entidades Estáticas e Dinâmicas:** A definição de quais entidades no processo de instanciação seriam estáticas ou dinâmicas foi realizado em conjunto com o professor Dr. Elvis Fusco coordenador do grupo de pesquisa ITIC (Inovação em Tecnologias Computacionais Digitais), professor Dr. Fábio Dacêncio Pereira orientador deste Trabalho de Conclusão de Curso e o autor deste trabalho. Logo, levou-se em consideração como dinâmico aquilo que tende a ser volátil num curto período de tempo e o restante foi estabelecido como dinâmico. Desse modo, a seguir observa-se uma tabela com o que foi definido:

Tabela 5.1: Tabela de entidades estáticas e dinâmicas

Entidade Estática	Entidade Dinâmica
CentroPrivadoPDI	Noticia
LaboratorioPDI	Projeto
Coworking	Evento
AmbienteInovacao	Pesquisador
Crowdfunding	Legislacao
Aceleradora	Edital
FundoAmparo	AreaInovacao
AgenciaFomento	
Investidor	
AgenciaInovacao	
Consultoria	
ArranjoProdutivoLocal	
EntidadeClasse	
InstituicaoEnsinoPesquisa	
OrgaoPublico	
EntidadeFomento	

EcossistemaPaulistaInovacao	
AmbienteFormalInovacao	
ParqueTecnologico	
CentroInovacaoTecnologico	
IncubadoraBaseTecnologica	
NucleoInovacaoTecnologica	
Empresa	
SistemaLocalInovacao	
EntidadeApoioInovacao	

- **Modificação do diagrama UML do InovaOnto:** A seguir segue uma lista com as mudanças e posteriormente a nova imagem do diagrama UML:
  - Adicionada a classe “Noticia”;
  - Foi criado propriedades (Entidade, Nome, Site) para a persistência semântica de classes estáticas;
  - Foi criado propriedades (Link, Domínio, Título, Descrição) para a persistência semântica de classes estáticas.

Inclusive, as modificações citadas anteriormente realizadas no diagrama UML da InovaOnto podem ser observadas no Anexo A.

- **Modificação do OWL do InovaOnto:** A seguir segue uma lista com as mudanças e posteriormente uma representação gráfica da Ontologia Conceitual do Ecossistema Paulista de Ambientes de Inovação:
  - Foi adicionado a entidade “Noticia”;
  - Implementou-se propriedades para todas as entidades, no entanto, as que serão instanciadas estaticamente receberam (Entidade, Nome e Site), já as dinâmicas (Link, Domínio, Título e Descrição)**.**

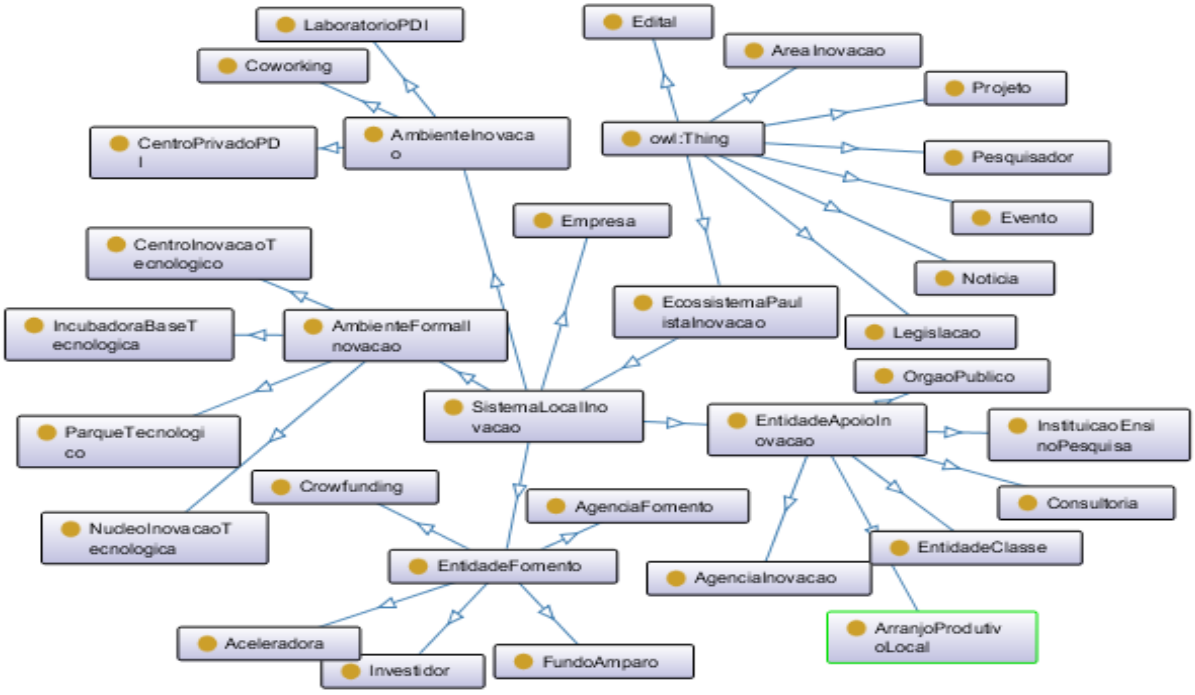


Figura 5.7: Ontologia alterada do InovaOnto – Fonte: Adaptada de Fusco (2017)

5.5 Classificação e persistência semântica

A classificação semântica deste Trabalho de Conclusão de Curso é realizada com o auxílio da ferramenta criada pela Apache denominada Jena, desse modo, neste subcapítulo será apresentado como é realizado o processo de instânciação estática, instanciação dinâmica, persistência no servidor de triplas FUSEKI e no banco de dados MySQL.

Dessa forma em seguida, é apresentado um fluxograma de como são realizados esses processos:

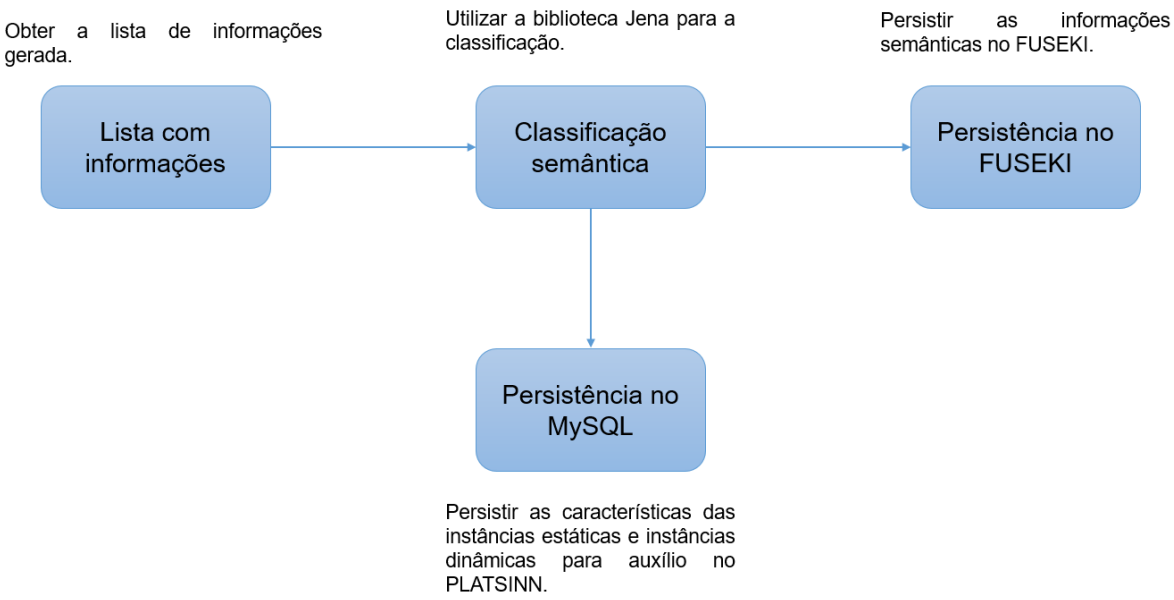


Figura 5.8: Fluxograma da classificação semântica e da persistência

Como mostrado na figura 5.8 o processo de classificação e persitência semântica é dividido em 4 partes denominadas: “Lista com informações”; “Classificação semântica”; “Persistência no FUSEKI”; “Persistência no MySQL”;

- **Lista informações:** Primeiramente, a plataforma receberá a lista de metadados de informações produzidas na seção 5.3;
- **Classificação com tecnologias da Web Semântica:**
    - **Instâncias estáticas:** Neste trabalho necessita-se a criação de instâncias estáticas para o cadastro dos ambientes informacionais digitais que serão recolhidas as informações de notícias, editais e eventos.

      Dessa forma, a biblioteca que auxilia a criação dessas instâncias é o Jena. Portanto, a seguir encontra-se uma imagem de um código em Java com a finalidade de implementar a categorização semântica estática:

```
1 OntClass intanciaEstaticaOnto = ontologia.getOntClass(tipes.getOnt_class_uri());
  DatatypeProperty nomeOnto = ontologia.getDatatypeProperty(tipes.getOnt_nome_uri());
  DatatypeProperty siteOnto = ontologia.getDatatypeProperty(tipes.getOnt_site_uri());
  DatatypeProperty entidade = ontologia.getDatatypeProperty(tipes.getOnt_entidade_uri());

2 //***************INSTANCIAR A ONTOLOGIA***************
  Individual ind = intanciaEstaticaOnto.createIndividual(baseURI+instanciaEstatica.getCod_estatico());
  ontologia.createIndividual(ind);
3 ind.addLiteral(nomeOnto, instanciaEstatica.getNome());
  ind.addLiteral(siteOnto, instanciaEstatica.getSite());
  ind.addLiteral(entidade, ent.obterEntidade(instanciaEstatica).getNome());
```

Figura 5.9: Exemplo de um código em Java para criar instâncias estáticas

A seguir é apresentado uma lista com a explicação de cada parte numerada presente na figura 5.9:

1. Parte do código responsável por referenciar as URIs da ontologia com atributos e objetos do Jena;
2. Nesta etapa, ao invocar o método “createIndividual()” a um “OntClass” é criado em memória a instância da ontologia, sendo que, é necessário dar um identificador a instância (URI base acrescida de um código de instância estática);
3. Nesta parte é relacionado as URIs que referenciamos anteriormente no passo 1 com suas devidas informações, ou seja, estamos adicionando propriedades a instância que foi gerada.

- **Instâncias dinâmicas:** A criação de instâncias estáticas é destinada a classificação de notícias, editais ou eventos, pois suas a cargas informacionais em ambientes digitais são muito elevadas.

  Desse modo, a seguir consta um exemplo de código de como é feito a implementação de instâncias dinâmicas utilizando o apache Jena:

Figura 5.10: Exemplo de um código em Java para criar instâncias dinâmicas

A seguir é apresentado uma lista com a explicação de cada parte numerada presente na figura 5.10:

1. Parte do código responsável por referenciar as URIs da ontologia com atributos e objetos do Jena;
2. Nesta étapa é criado a instância da ontologia, adicionadas suas propriedades através da relação com as URIs referenciadas anteriormente e seus respectivos conteúdos;
3. Nesta parte é recuperado uma lista com as instâncias estáticas persistidas na base de dados do FUSEKI e posteriormente é resgatado qual o código da instância estática que possui como domínio o mesmo das informações que estão sendo inseridas;
4. Com o código das instâncias estáticas que contém o domínio das informações é realizado uma relação de propriedade entre as instâncias estáticas e dinâmicas. Desse modo, o motor semântico consegue inferir quais informações

pertencem a entidades estáticas. Por exemplo: é possível realizar uma busca de quais as notícias que pertencem à Órgãos Públicos.

- **Persistência no FUSEKI:** Após a categorização semântica, o PLATSINN cria uma instância do modelo da ontologia existente no servidor com as novas informações.
- **Persistência no MySQL:** Após as informações serem classificadas e persistidas utilizando tecnologias da Web Semântica é preciso armazenar num banco de dados relacional alguns detalhes para que beneficie a usabilidade do cliente da Plataforma de apoio à Inovação.

Desse modo, foi desenvolvido um modelo de entidade relacionamento que pode ser observado a seguir:

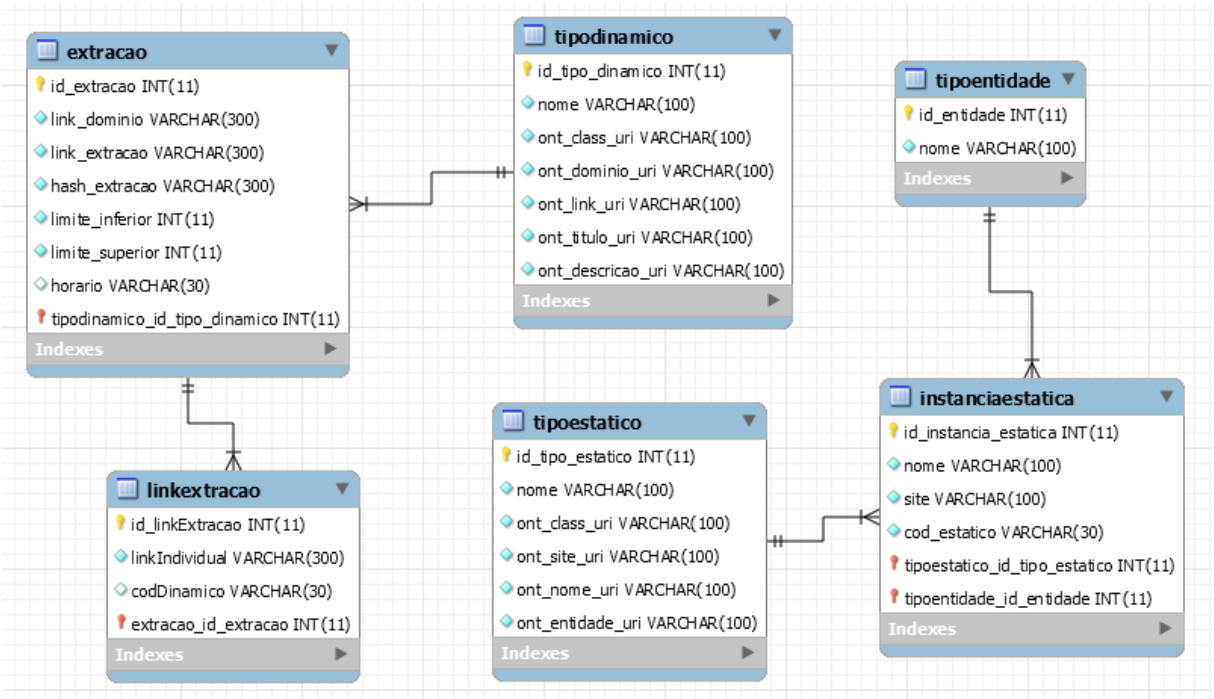


Figura 5.11: Modelo entidade relacionamento para apoio ao PLATSINN

Conforme a 5.11 é visível que a criação das entidades “tipoestatico” e “tipodinamico” apoiam o usuário produzir instâncias estáticas e dinâmicas, pois ele não necessita de um conhecimento aprofundado da ontologia InovaOnto. Inclusive,

alguns atributos dessas entidades são destinados ao preenchimento por parte do autor deste trabalho.

Além disso, exemplifca-se o constatado anteriormente da seguinte maneira: caso algum cliente queria criar na plataforma uma instância estática ele não precisa informar: “nome”, “ont_class_uri”, “ont_site_uri”, “ont_nome_uri”, “ont_entidade_uri”, pois já foram criados objetos tipoestatico previamente no banco de dados pelo desenvolvedor do PLATSIN. O desenvolvimento de objetos prévios no banco de dados relacional das classes: “tipoentidade”, “tipoestatico” e “tipodinamico” é uma realidade deste trabalho para apoiar o cliente na utilização da plataforma.

5.6 Classificação Cognitiva

A classificação cognitiva deste Trabalho de Conclusão de Curso é realizada através da integração com o IBM Watson. Logo, houve-se a junção de duas funcionalidades confeccionadas pela a IBM o “*Language Translator*” e o “*Natural Language Understanding*”, pois o mecanismo responsável por compreender a linguagem natural de textos escritos no idioma Português e retornar uma análise cognitiva não entrega todas as características necessárias para a Plataforma de Apoio à Inovação. Portanto, quando um usuário necessita de uma análise cognitiva de uma informação ela é traduzida para o idioma inglês através do “*Language Translator*” e a resposta desse serviço é transmitida ao “*Natural Language Understanding*”.

Primeiramente, para a utilização do IBM Watson é preciso criar uma conta neste endereço: “https://console.bluemix.net/” e posteriormente no painel do IBM Bluemix realizar o cadastro de utilização destas duas funcionalidades:

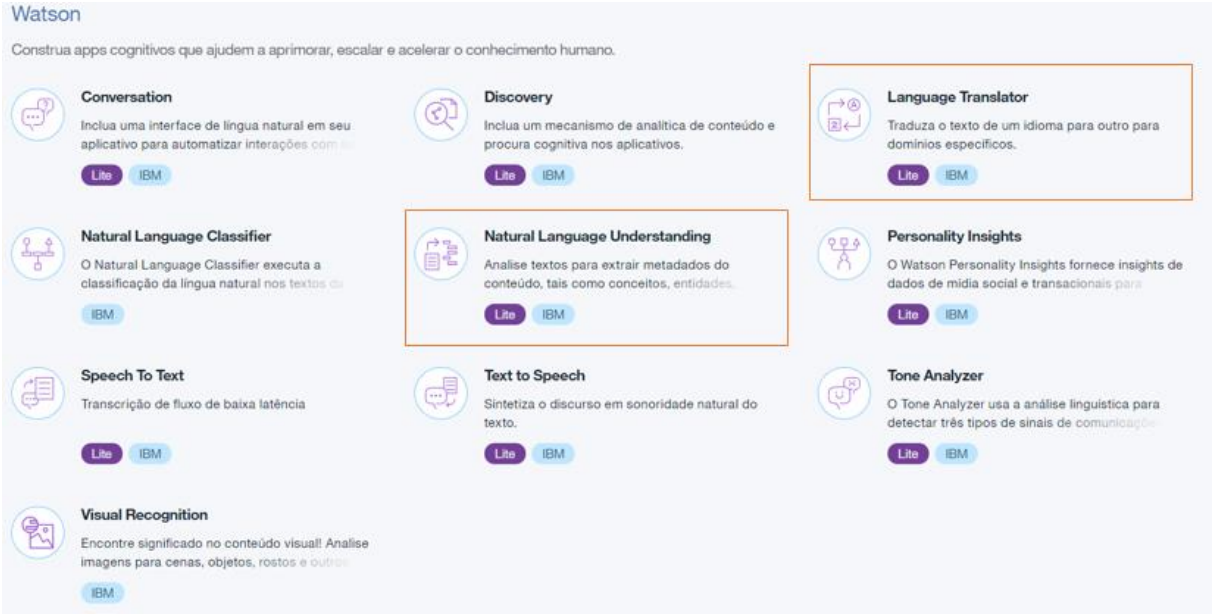


Figura 5.12: Catálogo de funcionalidades do IBM Watson – Fonte: Adaptado do catálogo do IBM Watson (https://console.bluemix.net/catalog/)

Após o cadastro nas duas funcioanalidades demonstradas na figura 5.12 é necessário a recuperação das credenciais de integração externa de ambas, no qual podem ser observadas ao acessar cada serviço e ir na aba denominada “Credenciais de serviço”. Feito isso, é possível realizar a tradução ou a análise cognitiva de um conteúdo na própria plataforma da IBM ou através da integração com alguma aplicação externa.

5.6.1 Language Translator

O serviço “*Language Translator*” do IBM Watson tem a capacidade de traduzir um texto de um idioma para outro e necessita de alguns parâmetros de entrada para que seja possível essa tradução. Portanto, a seguir encontra-se quais são os parâmetros de uma requisição ao serviço de tradução do Watson:

- URL da API (“https://gateway.watsonplatform.net/language-translator/api”);
- Credenciais (login e senha);
- O conteúdo que deve ser analizado;
- Em qual idioma se encontra o texto;
- Qual o idioma se deseja obter o conteúdo requisitado.

Desse modo, caso traduzirmos esta frase de Português do Brasil para Inglês dos Estados Unidos: “Plataforma de apoio à Inovação que tem a capacidade de extrair, classificar e recuperar informações semânticas e cognitivas”, utilizando o serviço de tradução do Watson obteremos uma resposta em formato JSON da seguinte maneira:


```
{
  "word_count": 19,
  "character_count": 123,
  "translations": [
    {
      "translation": "Platform support for Innovation that has the ability to extract, sort and retrieve information semantics and cognitive"
    }
  ]
}
```


Figura 5.13: Exemplo de JSON de tradução

Logo, abaixo é apresentado uma lista com a descrição dos atributos presentes no arquivo JSON da 5.13 conforme IBM Watson (2017):

- **world_count:** quantidade de palavras;
- **character_count:** quantidade de caracteres;
- **translation:** tradução do texto informado.

Para mais informações de como foi realizado esta implementação em Java vá até o apêndice 1 ou acesse a documentação da funcionalidade no site da IBM (https://www.ibm.com/watson/developercloud/language-translator/api/v2/) .

5.6.2 Natural Language Understand

O serviço “*Natural Language Understanding*” do IBM Watson possui a capacidade de analisar um texto e utilizando um mecanismo de compreensão de linguagem natural infere quais são as palavras-chave, conceitos, entidades, categorias, emoções, sentimentos e relações presentes. Dessa forma, a seguir encontra-se quais são os parâmetros de uma requisição ao serviço de interpretação de linguagem natural do Watson:

- URL da API (https://gateway.watsonplatform.net/natural-language-understanding/api);
- Credenciais (login e senha);

- O conteúdo que deve ser analizado;
- Qual a versão do “*Natural Language Undertanding*” deseja ser utilizada (VERSION_DATE_2017_02_27);
- Quais características que devem ser analizadas no conteúdo informado:
    - Conceitos envolvidos;
    - Categorias;
    - Entidades;
    - Emoções;
    - Palavras-chave;
    - Sentimento.

Desse modo, caso realizarmos uma análise cognitiva desta frase: “Plataforma de apoio à Inovação que tem a capacidade de extrair, classificar e recuperar informações semânticas e cognitivas”, onde requere-se categorias, palavras-chave, conceitos, emoções e sentimentos relacionados utilizando o serviço de compreensão de linguagem natural do Watson obteremos uma resposta em formato JSON com o conteúdo no idioma inglês da seguinte maneira:


```
"categories": [
  {
    "label": "\/science\/social science\/linguistics",
    "score": 0.402447
  },
  {
    "label": "\/technology and computing\/software\/databases",
    "score": 0.382067
  },
  {
    "label": "\/technology and computing",
    "score": 0.234043
  }
],
```


Figura 5.14: Categorias obtidas da análise cognitiva no formato JSON

A seguir é apresentado uma lista com a descrição dos atributos relevantes presentes no arquivo JSON da 5.14 conforme IBM Watson (2017):

- **label:** Nome hierárquico da categoria que o texto pode ser classificado;
- **score:** A pontuação da categorização, no qual varia de 0 até 1. Sendo que, quanto mais próximo do limite inferior há uma desconfiança na categorização,

outrora, quanto mais perto do limite superior existe uma confiança na categorização.


```
"keywords": [
  {
    "relevance": 0.950395,
    "text": "retrieve information semantics",
    "sentiment": {
      "score": -0.321718
    }
  },
  {
    "relevance": 0.633961,
    "text": "Platform support",
    "sentiment": {
      "score": 0.341193
    }
  },
  {
    "relevance": 0.519532,
    "text": "sort",
    "sentiment": {
      "score": -0.321718
    }
  },
  {
    "relevance": 0.49438,
    "text": "Innovation",
    "sentiment": {
      "score": 0.341193
    }
  },
  {
    "relevance": 0.470042,
    "text": "ability",
    "sentiment": {
      "score": 0.341193
    }
  }
],
```


Figura 5.15: Palavras-chave obtidas da análise cognitiva no formato JSON

A seguir é apresentado uma lista com a descrição dos atributos relevantes presentes no arquivo JSON da 5.15 conforme IBM Watson (2017):

- **text:** Nome da palavra-chave;
- **relevance:** A relevância da palavra-chave, no qual varia de 0 até 1. Sendo que, quanto mais próximo do limite inferior é ela se torna mais irrelevante, outrora, quanto mais perto do limite superior ela se torna mais relevante.

- **sentiment:** A pontuação sentimental para a palavra-chave, onde pontuações negativas indicam sentimentos negativos e pontuações positivas apontam sentimentos positivos.


```
"concepts": [
  {
    "text": "Maximum likelihood",
    "relevance": 0.852274,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Maximum_likelihood"
  },
  {
    "text": "Francis Ysidro Edgeworth",
    "relevance": 0.823906,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Francis_Ysidro_Edgeworth"
  },
  {
    "text": "Perception",
    "relevance": 0.813686,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Perception"
  },
  {
    "text": "Language",
    "relevance": 0.788978,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Language"
  },
  {
    "text": "Charles Sanders Peirce",
    "relevance": 0.701973,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Charles_Sanders_Peirce"
  },
  {
    "text": "Information processing",
    "relevance": 0.678971,
    "dbpedia_resource": "http:\/\/dbpedia.org\/resource\/Information_processing"
  }
],
```


Figura 5.16: Conceitos relacionados obtidos da análise cognitiva no formato JSON

Em seguinda é apresentado uma lista com o significado de cada característica contida no JSON da 5.16 conforme IBM Watson (2017):

- **text:** Qual o nome do conceito relacionado ao conteúdo informado;
- **relevance:** A relevância do conceito apresentado, onde varia de 0 até 1. Inclusive, quanto mais próximo de 0 o conceito tende a ser irrelevante, outrora, quanto mais próximo de 1 caminha para uma relevância definitiva;
- **dbpedia_resource:** URL no portal do DBpedia para uma pesquisa mais aprofundada a respeito do conceito.

```
"emotion": {
  "document": {
    "emotion": {
      "anger": 0.043154,
      "disgust": 0.088304,
      "fear": 0.034827,
      "joy": 0.264291,
      "sadness": 0.078887
    }
  }
},
```


Figura 5.17: Emoções obtidas da análise cognitiva no formato JSON

Conforme IBM Watson (2017) os atributos presentes na 5.17, são: **anger** (raiva); **disgust** (desgosto); **fear** (medo); **joy** (alegria); **sadness** (tristeza). Inclusive, o valor em frente ao atributo significa que quanto mais próximo de 0 o texto não transmite a emoção, outrora quanto mais próximo de 1 o conteúdo definitivamente transmite.


```
"sentiment": {
  "document": {
    "score": 0.765067
  }
}
}
```


Figura 5.18: Sentimento obtido da análise cognitiva no formato JSON

Segundo IBM Watson (2017), a característica **sentiment** (sentimento) presente na figura 5.18 é obtida a partir de uma análise em todo o texto informado, sendo que o **score** (pontuação) resultante varia de -1 (sentimento negativo) até +1 (sentimento positivo), além disso, caso for 0 significa que o conteúdo do texto é neutro.

5.7 Recuperação da Informação

A recuperação da informação na Plataforma de apoio à Inovação é feita por meio de consultas SPARQL no servidor de base de dados em RDF denominado FUSEKI. Dessa forma, há duas maneiras para que ela seja realizada:

- Integração com aplicações externas;
- Recuperação na página dataset.html do próprio servidor;

5.7.1 Integração com aplicações externas

Para a realização de uma consulta no FUSEKI através de uma integração com um serviço externo é preciso efetuar uma requisição HTTP, portanto, a seguir consta um exemplo de como um cliente consultaria informações no servidor:

- **URL:** Endereço da sua base de dados acrescido com “/sparql” (http://localhost:3030/inovaOnto/sparql).
- **Forma de Requisição:** GET ou POST
- **Parâmetros:**
    - **query:** Código SPARQL;
    - **format:** Formato da resposta (JSON, XML, etc.).

Dessa maneira, para a consulta de uma notícia do domínio (“www.umbco23.com.br”) há como resposta o seguinte JSON:

```
{"head": {"vars": ["descricao","link"]},
  "results": {
    "bindings": [
      {
        "descricao": {
          "type": "literal",
          "value": "Alternar a navegaçãoSala de reuniãoNegócios,
atendimentos e treinamentos.Nossa sala de reunião comporta até 8
pessoas confortavelmente com notebook e internet. Ou até 15
pessoas no formato auditório. Deixamos servido café e água a
vontade. A sala é equipada com ar-condicionado, projetor e uma
tela de vidro usada como lousa.Nossa sala de reunião comporta até
8 pessoas confortavelmente com notebook e internet. Ou até 15
pessoas no formato auditório. Deixamos servido café e água a
vontade. A sala é equipada com ar-condicionado, projetor e uma
tela de vidro usada como lousa.1 horaR$ 40,00DiáriaValores válidos
para uso de segunda a sexta das 8hs às 19hs. Para outros dias e
horários, solicite um orçamento. recepcao@umbco23.com.brValores
válidos para uso de segunda a sexta das 8hs às 19hs. Para outros
dias e horários, solicite um orçamento. recepcao@umbco23.com.brSeu
nome (obrigatório)Seu e-mail (obrigatório)Seu telefone (obrigatóri
o)Dia e hora da reservaPS: Aguarde nosso contato para confirmação.
Pagamento antecipado ou na hora em dinheiro.Melhor Coworking de
CampinasNomad CapitalistGo to FacebookGo to TwitterGo to LinkedinU
mb.co23 Coworking 2017Alternar a navegação"
        },
        "link": {
          "type": "literal",
          "value": "http://umbco23.com.br/sala-de-reuniao/"
        }}]}}
```


Figura 5.19: Notícia recuperada do domínio umbco23 em formato JSON

Abaixo é apresentado uma lista com a descrição das características relevantes presentes no arquivo JSON resultante da requisição de filtro na base de dados:

- **vars:** Neste atributo consta quais são as variáveis envolvidas na resposta (descricao, link);
- **results > bindings:** Retorna uma lista com todas as informações das notícias resultantes da consulta SPARQL;
- **descricao > value:** Neste campo é indicado qual é a descrição da notícia;
- **link > value:** Atributo que tem a responsabilidade de conter o endereço da notícia pesquisada.

5.7.2 Recuperação na página do servidor

Para a recuperação das informações na página do FUSEKI é necessário acessar o endereço (http://localhost:3030/dataset.html) e posteriormente será possível observar a seguinte página:

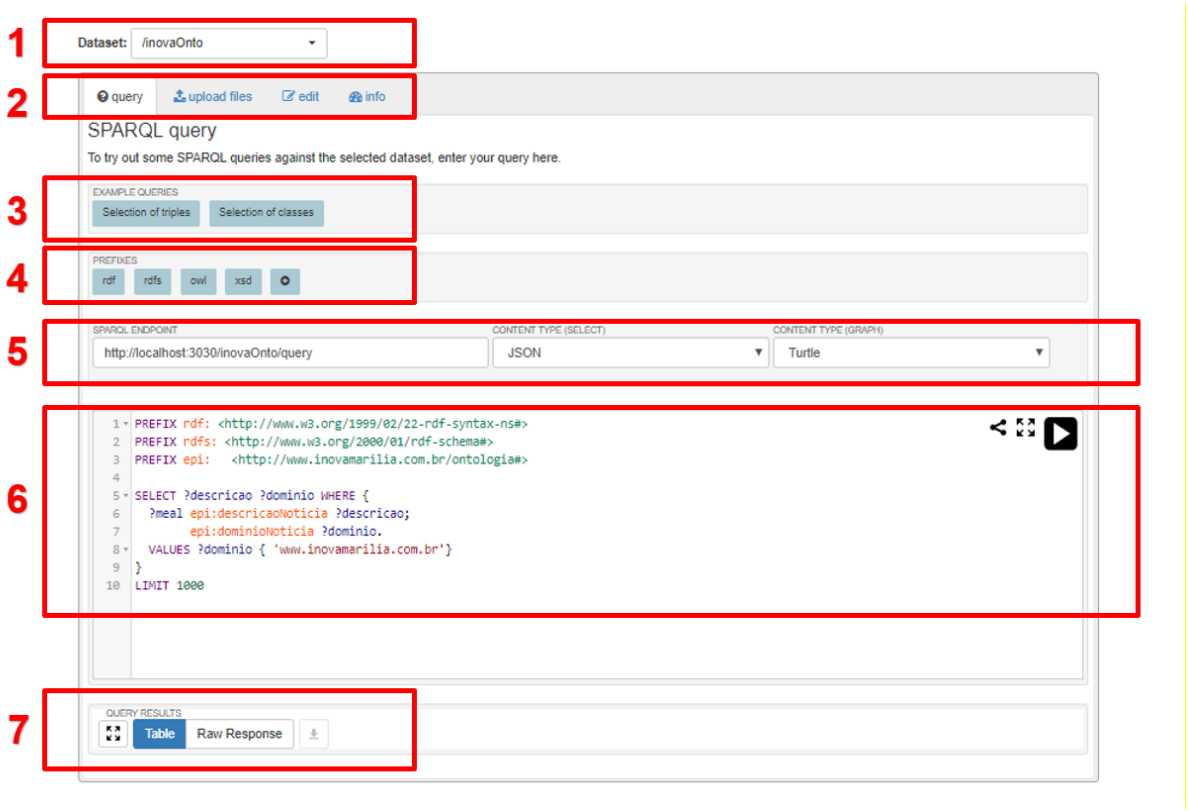


Figura 5.20: Tela da base de dados do FUSEKI

Conforme observado na 5.20 o autor deste trabalho dividiu a página em 7 seções para que seja possível uma melhor explicação das possibilidades de configuração que o servidor de persistência de triplas permite ao usuário no momento da recuperação da informação, logo segue a explanação:

1. Nesta seção é possível o usuário selecionar em qual base de dados ele deseja realizar uma consulta;
2. Neste seguimento, selecione respectivamente entre: filtrar informações; carregar arquivos rdf; editar a base de dados selecionada e obter informações de integração com serviços externos;

3. Exemplo de códigos SPARQL para filtrar todas as triplas ou somente as classes;
4. Possibilidade de escolher entre os prefixos que serão utilizados na consulta;
5. Informe respectivamete: qual a URL da requisição, o tipo de conteúdo para a seleção das informações e o tipo de dado persistido que se deseja consultar no FUSEKI;
6. O código SPARQL para a busca;
7. Como o usuário deseja que o servidor exiba os resultados se é em tabela ou em JSON, CSV etc.

5.8 API

A API criada para a Plataforma de Apoio à Inovação foi implementada com a ideia de integrar a plataforma com outras aplicações, como: aplicativos móveis, sistemas desktop, sistemas web, etc. Além disso, utilizou-se a linguagem de programação JAVA e o servidor de aplicação open source denominado GlassFish. Desse modo, suas funcionalidades vão desde a criação, exclusão e requisição de informações e seu tráfego é feito no formato JSON, pois ele garante a interoperabilidade entre diferentes linguagens de programação.

5.8.1 Cadastrar

Primeiramente, caso o usuário deseje cadastrar alguma informação na plataforma ele precisa informar qual é qual é a URL da ação desejada na API, ou seja a URL da ação, quais são os parâmetros e qual a forma de requisição (GET). Portanto, a seguir consta um exemplo de como um usuário cadastraria uma extração no servidor:

- **URL:**http://localhost:51202/ExtratorSemanticoCognitivo/CadastrarExtracaoServlet
- **Forma de Requisição:** GET
- **Parâmetros:**
    - **link:** http://www.inovamarilia.com.br/category/noticias/

- **tipo_dinamico:** noticia
- **profundidade:** 1

Dessa maneira, a URL da requição à API ficaria: (http://localhost:51202/ExtratorSemanticoCognitivo/CadastrarExtracaoServlet?link=http://www.inovamarilia.com.br/category/noticias/&tipo_dinamico=noticia&profundidade=1).

Logo, resultado deste requerimento é em formato JSON pode ser *true* (verdadeiro) para o sucesso da realização da obtenção da informação líquida pelo robô de busca ou *false* (false) para o insucesso da extração dos metadados e dados.

5.8.2 Deletar

Caso o cliente deseje deletar alguma informação na plataforma é necessário comunicar qual é a URL da ação desejada na API, quais são os parâmetros e qual a forma de requisição (GET). Portanto, em seguida é demontrado como um usuário deletaria as informações adquiridas de uma extração no servidor de aplicação:

- **URL:**http://localhost:51202/ExtratorSemanticoCognitivo/DeletarExtracaoServlet
- **Forma de Requisição:** GET
- **Parâmetros:**
  - **extracao:** www.inovamarilia.com.br/category/noticias/

Desse jeito, a URL da requição à API ficaria: (http://localhost:51202/ExtratorSemanticoCognitivo/DeletarExtracaoServlet?extracao=www.inovamarilia.com.br/category/noticias/).

Logo, resultado deste requerimento é em formato JSON e pode ser *true* (verdadeiro) para o sucesso da exclusão das informações ou *false* (false) para o insucesso.

5.8.3 Listar

Se o usuário desejar obter informações específicas persistidas pelo servidor de aplicação é necessário que ele comunique qual é o endereço da função de

listagem desejada na API e sua forma de requisição (GET). Então, a seguir é demonstrado como um cliente obteria uma lista de dados persistidos na Plataforma de Apoio à Inovação:

- **URL:**http://localhost:51202/ExtratorSemanticoCognitivo/ListarExtracaoServlet
- **Forma de Requisição:** GET

Desse modo, a URL da requição à API ficaria: (http://localhost:51202/ExtratorSemanticoCognitivo/ListarExtracaoServlet).

O resultado obtido ao requerer essas informações é em formato JSON e pode ser observado a seguir:


```
[
  {
    "id": 45,
    "linkDominio": "www.parquebtu.org.br",
    "linkExtracao": "http:\/\/www.parquebtu.org.br\/editais-finep\/editais2",
    "limiteInferior": 321,
    "limiteSuperior": 331,
    "hashExtracao": "5C55638FD31DEE651874206A897859D7AFBD7B4BFD56FF75EA5315A7BB41B198",
    "horario": "24\/10\/2017 01:30:51",
    "fk_id_tipo": 3
  },
  {
    "id": 43,
    "linkDominio": "www.baita.ac",
    "linkExtracao": "http:\/\/www.baita.ac\/category\/news\/",
    "limiteInferior": 257,
    "limiteSuperior": 312,
    "hashExtracao": "02E7098F3736C714BE4E3979136CCC0E4907867CBDAF30874B83ABAF2477E660",
    "horario": "24\/10\/2017 01:16:39",
    "fk_id_tipo": 1
  },
  {
    "id": 68,
    "linkDominio": "www.inovamarilia.com.br",
    "linkExtracao": "http:\/\/www.inovamarilia.com.br\/category\/editais\/",
    "limiteInferior": 563,
    "limiteSuperior": 598,
    "hashExtracao": "BE33A99AB81178CCEADD4A74149B0554A47D7E159423F8FAC07A40DF955EDF5F",
    "horario": "09\/11\/2017 00:14:27",
    "fk_id_tipo": 3
  }
]
```


Figura 5.21: Extrações no formato JSON

Abaixo é apresentado uma lista com a descrição dos atributos presentes no arquivo JSON resultante da requisição de listagem de extrações à API:

- **id:** Identificador;
- **linkDominio:** Domínio que foi realizado a extração;

- **linkExtracao:** Qual a seção do site informado para a realização de obtenção de informação líquida.
- **limiteInferior:** Qual o identificador mínimo das informações que foi persistida no modelo ontológico;
- **limiteSuperior:** Qual o identificador máximo das informações que foi persistida no modelo ontológico;
- **hashExtracao:** Hash dos metadados e dados obtidos com a extração para verificação de duplicidade informacional;
- **horario:** O horário que foi realizado a extração;
- **fk_id_tipo:** Qual o identificador do tipo de extração (Notícia, Edital, Evento, etc.) que está sendo realizado.

5.8.4 Funções da API

A API desenvolvida para a Plataforma de Apoio à Inovação realiza diversas funções que vão desde cadastrar e deletar uma extração até a obtenção de uma análise cognitiva em uma notícia, edital ou evento através do IBM Watson, portanto em seguida pode-se conferir todas as funções implementadas à API neste Trabalho de Conclusão de Curso com suas respectivas descrições, formas de requisição, parâmetros, tipo de saída e exemplos de requisição:

Tabela 5.2: Tabela com as funções da API

Ação	Descrição	Parâmetros	Saída
/ListarTipoEstaticoServlet	Retorna uma lista de tipos estáticos.		JSON
/ListarTipoDinamicoServlet	Retorna uma lista de tipos dinâmicos.		JSON
/ListarInstanciaEstaticaServlet	Retorna uma lista de instâncias estáticas.		JSON
/ListarInstanciaDinamicaServlet	Retorna uma lista de instâncias		JSON

	dinâmicas.		
/ListarExtracaoServlet	Retorna uma lista de extrações.		JSON
/ListarEntidadeServlet	Retorna uma lista de entidades.		JSON
/DeletarInstanciaEstaticaServlet	Deleta uma instância estática.	codigo	JSON
/DeletarExtracaoServlet	Deleta uma extração.	dominio	JSON
/CadastrarInstanciaEstaticaServlet	Cadastra uma instância estática.	nome, site, id_entidade, id_tipo_estatico	JSON
/CadastrarExtracaoServlet	Cadastra uma extração.	Link, tipo_dinâmico, profundidade	JSON
/ObterAnaliseCognitiva	Obtém uma análise cognitiva de um conteúdo utilizando o IBM Watson!	texto	JSON

Capítulo 6

CONCLUSÃO

Este capítulo tem a finalidade de relatar quais foram os resultados obtidos com a Plataforma de Apoio à Inovação denominada PLATSINN, as contribuições e limitações deste Trabalho de Conclusão de Curso e quais são as oportunidades para trabalhos futuros.

6.1 Validação do robô de busca

Primeiramente, este subcapítulo tem a função de validar o robô de busca, sendo que foi realizado uma pesquisa com seis usuários especialistas da área de Ciência da Computação sobre a congruência entre os resultados obtidos com o robô de busca e os metadados e dados que eles encontraram em páginas que foram extraídas. Inclusive, o resultado foi publicado no 14º International Conference on Information Systems and Technology Management (CONTECSI).

6.1.1 Metodologia de validação do agente

- Criar formulário de avaliação da ferramenta;
- Definir os espaços informacionais que serão utilizados para a extração;

- Definir os usuários especialistas que serão os avaliadores;
- Extrair os metadados e dados das fontes informacionais selecionadas;
- Explicar ao usuário especialista a forma de avaliação da ferramenta e demonstrá-lo manualmente metadados e dados em uma página para reforçar o seu conhecimento;
- Pedir ao usuário para selecionar uma URL aleatória;
- Pedir ao usuário para contar quantos metadados e dados ele acha ao acessar o endereço Web selecionado;
- Analisar a quantidade de metadados e dados que a ferramenta retornou;
- Comparar quantitativamente os resultados obtidos pelo usuário com o da ferramenta;
- Obter gráficos com a comparação dos resultados obtidos.

6.1.2 Formulário de validação do robô de busca

O formulário de validação do agente de extração de metadados e dados foi criado com auxílio da ferramenta Google Forms, logo segue a representação do guia na imagem a seguir:

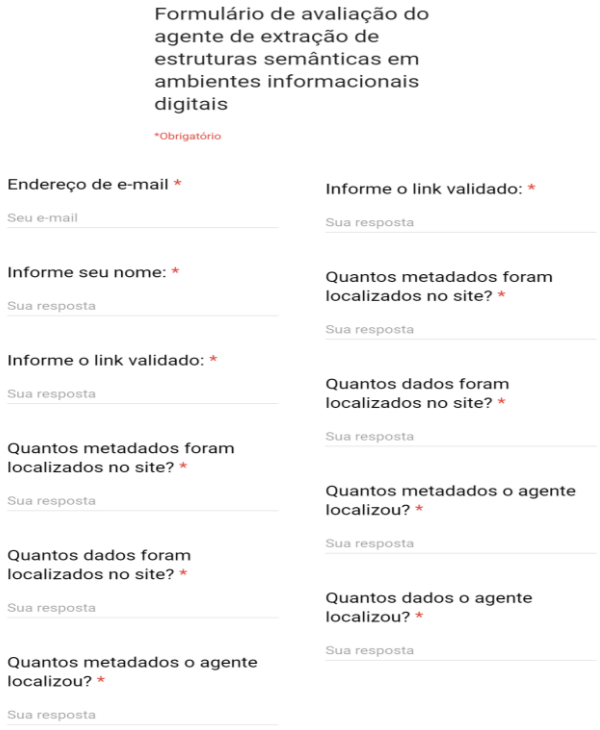


Figura 6.1: Formulário de validação do agente – Fonte: Adaptado do Google Forms (https://goo.gl/forms/xGHQAEnDUfMFWMNJ2)

Conforme mostrado anteriormente, o formulário de avaliação do mecanismo de extração foi elaborado para que houvesse uma congruência entre os resultados informados pelos usuários especialistas e os que o agente entregou como resposta.

6.1.3 Resultados da validação

Como teste para verificar se o agente está extraindo os metadados e dados dos ambientes informacionais digitais corretamente houvesse a necessidade de escolher dois ambientes, logo foram selecionados o site da FAPESP (www.fapesp.br) e o Inova Marília (www.inovamarilia.com.br).

Dessa forma, os entrevistados escolheram links aleatoriamente e responderam o questionário com o auxílio de um avaliador juntamente com o sistema em execução.

Figura 6.2: Gráfico da relação entres os resultados extraídos no site da FAPESP

Figura 6.3: Gráfico da relação entres os resultados extraídos e os observados no site do Inova Marília

Portanto, com base nos gráficos obtidos do primeiro cenário o agente encontrou 13 metadados e os entrevistados 16, já em relação aos dados o mecanismo encontrou 113 e os usuários especialistas 128, desse modo, realize-se dois cálculos de porcentagem sobre os dois montantes de metadados e dados e uma média aritmética entre os resultantes é possível obter que a acurácia do agente para esse ambiente informacional é de 84,766% com base nos endereços eletrônicos validados.

Além disso, conforme as representações gráficas dos resultados obtidos no segundo espaço informacional digital o agente encontrou 10 metadados e os usuários também 10, já em relação aos dados o robô achou 165 e os usuários especialistas 148, portanto, ao realizar uma conta de porcentagem simples entre os montantes de metadados e dados e com os resultados aplicar uma média aritmética é possível obter o valor de 94,85%.

Conclui-se que o agente de extração e identificação de estruturas semânticas de ambientes informacionais digitais obteve uma exatidão média de 89,81%.

6.2 Considerações finais

Este trabalho tem como característica o desenvolvimento de uma Plataforma de Apoio à Inovação que tenha a capacidade de extrair, classificar e recuperar informações semânticas e cognitivas de ambientes informacionais digitais em prol de ajudar a tomada de decisão dos atores de inovação (governo, universidades e empresas).

Para o desenvolvimento do mecanismo de obtenção de informação líquida foi utilizado a linguagem de programação JAVA e o auxílio da biblioteca de manipulação de HTML Jsoup, sendo que os metadados e dados capturados estavam disponibilizados em tags predefinidas.

A transformação dos metados e dados de páginas Web em informação é realizada através de mecanismos computacionais capazes de obter o título e descrição contidos numa lista de objetos de “MDado”.

A alteração na ontologia inovaOnto teve o intuito de aderir propriedades inexistentes à ela para que seja possível integrá-la ao robô de busca criado neste trabalho. Além disso, o processo de criação de atributos para essa ontologia foi realizado com o auxílio da ferramenta Protége.

A classificação e persistência semântica das informações foi feita com o auxílio da biblioteca Jena e do servidor FUSEKI. Onde, as informações depois de categorizadas em instâncias estáticas ou dinâmicas são transmitidas por intermédio de uma conexão HTTP com o servidor de base da dados.

A recuperação da informação pode ser feita através da interface Web do PLATSINN ou pelo painel de controle do FUSEKI, inclusive, neste último é possível a exportação dos dados para formato JSON, CSV, entre outros.

A API foi criada com o intuito de permitir que outras aplicações futuramente sejam integradas ao PLATSINN, por isso varias funcionalidades do sistema tiveram uma conexão com o meio externo. Inclusive, esta foi desenvolvida com o apoio do servidor de aplicação GlassFish e Servlets.

A interface WEB deste Trabalho de Conclusão de Curso foi criada focada na usabilidade de um usuário do PLATSINN, onde o mesmo não necessita de conhecimentos sobre a InovaOnto, pois ela abstrai as URIs relacionados no processo de extração, exclusão e atualização da base de dados da plataforma.

Além disso, este trabalho seguiu os nove passos da metodologia de desenvolvimento que são: levantamento bibliográfico e pesquisa por trabalhos correlatos, estudo do cenário, desenvolvimento do robô de extração, mapeamento das informações captadas com o robô de busca, persitência desses dados coletados, recuperação da informação, análise cognitiva por intermédio do IBM Watson, criação de uma API para integrar o serviço do PLATSINN com aplicações externas e desenvolvimento de uma interface Web para a Plataforma de Apoio à Inovação.

O autor deste trabalho agradece ao CNPQ e a FAPESP pelas bolsas de iniciação científica confiadas a ele no decorrer destes quatro anos nos seguintes processos(Apoio CNPQ: 129317/2014-4, 146792/2015-7, 118142/2016-0; Apoio FAPESP: 2016/13025-0), no qual tiveram um papel crucial para o seu desenvolvimento acadêmico e como pessoa.

Além disso, gratifica-se o intercâmbio para a Europa fomentado pelo Santander Universidades e o apoio acadêmico do laboratório de pesquisa COMPSI no decorrer destes quatro anos de graduação.

Dessa forma, conclui-se que o objetivo deste trabalho que é desenvolver uma Plataforma de Apoio à Inovação que extrai, classifica e recupera informações semânticas e cognitivas produzidas pelos atores de inovação (governo, empresas e universidades) em um espaço informacional delimitado à Parques Tecnológicos, Centros de Inovação Tecnológica, entre outros credenciados ao Sistema Paulista de Ambientes de Inovação (SPAI) foi alcançado e com o desenvolvimento deste projeto foi possível por em prática o conteúdo assimilado dentro e fora da sala de aula.

6.3 Contribuições

As contribuições deste trabalho são destinadas desde o desenvolvimento da plataforma até a documentação de como foi realizado os processos de identifcação de metadados e dados, extração desses dados semi-estruturados, classificação e persistência semântica das informações obtidas, recuperação dessas informações e análise cognitiva de notícias, editais e eventos através do IBM Watson.

6.4 Limitações

Ao desenvolver a Plataforma de Apoio à Inovação (PLATSINN) o autor deste trabalho se deparou com as seguintes limitações:

- O mecanismo de extração de informações se delimita a seções de ambientes informacionas (notícias, editais ou eventos);
- A extração de metadados e dados é delimitada a tags específicas, logo, se o site não seguir o padrão convencional de disponibilização de conteúdo o robô de busca não obtera êxito em sua tarefa;
- Não é obtido informações de editais em formatos adversos de texto;
- A licença do IBM Watson utilizada tem limite de requisição, onde para o “Language Translator” a cota mensal é de um milhão de caracteres traduzidos, já para a “Natural Language Understand” são mil requisições mensais;
- Não é realizado um controle de usuários que acessam a plataforma, portanto se for disponibilizada em domínio público as funcionalidades de inclusão e exclusão deverão ser delimitadas.

6.5 Trabalhos Futuros

Neste subcapítulo consta quais são os próximos passos para o autor deste trabalho ou outros alunos e desenvolvedores continuarem este projeto.

6.5.1 Da plataforma

O projeto de desenvolvimento da Plataforma de Apoio à Inovação concluiu seu objetivo que é extrair, classificar e recuperar informações semânticas e cognitivas, logo há algumas funcionalidades interessantes para serem implementadas futuramente, como:

- Utilização de inteligência artificial e aprendizado de máquina ao mecanismo de extração para que seja possível a captação de informações sobre editais

em formatos adversos (PDF, DOCX, imagem, etc.) presentes nas seções dos sites;

- Criar mecanosmos de extração automática no PLATSINN, onde o usuário cadastro um horário para que todos os dias o sistema verifique se o site atualizou seu conteúdo;
- Disponibilizar o PLATSINN em domínio público, desse modo, a comunidade científica consegue agregar conhecimento ao projeto;
- Realizar validações na InovaOnto diferentes das utilizadas por Fusco (2017);
- Criar mecanismos de cadastro de usuários, login no sistema e controle de seção do cliente na plataforma, pois caso o projeto for disponível para a comunidade não serão todos os usuários capazes de manipular a base de dados do projeto;
- Definir filtros para o PLATSINN adversos aos existentes na plataforma, como: recuperação de notícias, editais e eventos por periodicidade e autoria.
- Desenvolver mecanismos capazes de cruzar informações, por exemplo: ao realizar a extração da seção de pesquisadores do site da FAPESP, extração de eventos de diversas fontes informacionais e cadastramento instâncias estáticas de empresas privadas. Logo, será possível através de uma tela obter quais são os pesquisadores relacionados a FAPESP que participaram de um evento numa determinada data que possuem projetos vinculados a empresas privadas cadastradas na base de dados.
- Explorar outros serviços cognitivos do IBM Watson, como: Discovery e Chat bot;
- Obter uma licença do IBM Watson que não possua limites de requisições, logo proporcionará maior liberdade para manipulação dos serviços da IBM e integração de novos módulos ao PLATSINN.

6.5.2 Da divulgação científica

A publicação dos resultados e desenvolvimento deste Trabalho de Conclusão de Curso pode ser realizada em eventos gerais, eventos específicos de divulgação científica e em Journals, como: Revista do IEEE América Latina; IoTBDS 2018; WorldCIST 2018; 15º CONTECSI.

6.6 Produção Bibliográfica

Neste subcapítulo é retrato quais são as informações acadêmicas relevantes do autor deste trabalho adquiridas no período do desenvolvimento deste projeto.

6.6.1 Bolsas de fomento a pesquisa e cultura

Bolsista de Iniciação em Desenvolvimento Tecnológico e Inovação - PIBITI; Projeto: Plataforma de Inteligência de Negócios Baseada em Estruturas Informacionais Semânticas: Modelo Computacional e Informacional de Apoio aos Ambientes de Inovação; Orientador: Elvis Fusco.

Bolsista de Iniciação Científica FAPESP; Projeto: Arquitetura de Apoio a Processos de Inovação Baseada em Estruturas Informacionais Semânticas; Orientador: Elvis Fusco.

Bolsista do Santander Universidades através do Programa de Bolsas Ibero-Americanas, onde visitei Lisboa, Munique, Berlim e Paris no período de 27/06/2017 até 11/07/2017.

6.6.2 Trabalhos completos publicados em anais de congressos

COSTA, T. A. G.; FUSCO, E. ; MUCHERONI, M. L. ; PEREIRA, F. D. ; CONEGLIAN, C. S. ; ORDONEZ, E. D. M. . **AGENTE DE EXTRAÇÃO E IDENTIFICAÇÃO DE ESTRUTURAS SEMÂNTICAS EM AMBIENTES INFORMACIONAIS DIGITAIS**. In: 14th CONTECSI International Conference on Information Systems and Technology Management, 2017, São Paulo. 14th CONTECSI International Conference on Information Systems and Technology Management, 2017. p. 5133-5150.

6.6.3 Resumos publicados em anais de congressos

COSTA, T. A. G.; PEREIRA, F. D. . **ARQUITETURA DE APOIO A PROCESSOS E AMBIENTES DE INOVAÇÃO BASEADA EM ESTRUTURAS INFORMACIONAIS SEMÂNTICAS E COGNITIVAS**. In: VII CONGRESSO DE PESQUISA CIENTÍFICA: INOVAÇÃO, SUSTENTABILIDADE, ÉTICA E CIDADANIA, 2017, Marília. CADERNO DE RESUMOS DO VII CONGRESSO DE PESQUISA CIENTÍFICA: INOVAÇÃO, SUSTENTABILIDADE, ÉTICA E CIDADANIA, 2017. p. 67-67.

COSTA, T. A. G.; PEREIRA, F. D. . **PLATAFORMA DE INTELIGÊNCIA DE NEGÓCIOS BASEADA EM ESTRUTURAS INFORMACIONAIS SEMÂNTICAS: MODELO COMPUTACIONAL E INFORMACIONAL DE APOIO AOS AMBIENTES DE INOVAÇÃO DO ESTADO DE SÃO PAULO**. In: VI CONGRESSO DE PESQUISA CIENTÍFICA: INOVAÇÃO, SUSTENTABILIDADE, ÉTICA E CIDADANIA, 2016, Marília. CADERNO DE RESUMOS DO VI CONGRESSO DE PESQUISA CIENTÍFICA: INOVAÇÃO, SUSTENTABILIDADE, ÉTICA E CIDADANIA, 2016. p. 51-51.

6.6.4 Apresentação de Trabalho

COSTA, T. A. G.; FUSCO, E. ; MUCHERONI, M. L. ; PEREIRA, F. D. ; CONEGLIAN, C. S. ; ORDONEZ, E. D. M. . **AGENTE DE EXTRAÇÃO E IDENTIFICAÇÃO DE ESTRUTURAS SEMÂNTICAS EM AMBIENTES INFORMACIONAIS DIGITAIS**. 2017. In:

COSTA, T. A. G.; PEREIRA, F. D. . **ARQUITETURA DE APOIO A PROCESSOS E AMBIENTES DE INOVAÇÃO BASEADA EM ESTRUTURAS INFORMACIONAIS SEMÂNTICAS E COGNITIVAS**. 2017.

COSTA, T. A. G.; PEREIRA, F. D. . **PLATAFORMA DE INTELIGÊNCIA DE NEGÓCIOS BASEADA EM ESTRUTURAS INFORMACIONAIS SEMÂNTICAS: MODELO COMPUTACIONAL E INFORMACIONAL DE APOIO AOS AMBIENTES DE INOVAÇÃO DO ESTADO DE SÃO PAULO**. 2016.

REFERÊNCIAS BIBLIOGRÁFICAS

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

BRASCHER, M. **WEB SEMÂNTICA**, 2007. Disponível em: http://www.stf.jus.br/arquivo/sijed/16.pdf. Acesso em: 09 de junho de 2017.

BREITMAN, K. K., **Web Semântica: A Internet do Futuro**, Rio de Janeiro: LTC, 2005.

CONEGLIAN, C. S. **Agente Semântico de Extração Informacional no Contexto de Big Data**. Trabalho de Conclusão de Curso para o grau de Bacharel em Ciência da Computação no Centro Universitário Eurípedes de Marília (UNIVEM), Marília, 2014.

COSTA, T. A. G.; FUSCO, E. ; MUCHERONI, M. L. ; PEREIRA, F. D.; CONEGLIAN, C. S. ; ORDONEZ, E. D. M. . **AGENTE DE EXTRAÇÃO E IDENTIFCAÇÃO DE ESTRUTURAS SEMÂNTICAS EM AMBIENTES INFORMACIONAIS DIGITAIS**. In: 14th CONTECSI International Conference on Information Systems and Technology Management, 2017, São Paulo. 14th CONTECSI International Conference on Information Systems and Technology Management, 2017. p. 5133-5150.

DIAS, T. D.; SANTOS, N. **Web Semântica: Conceitos Básicos e Tecnologias Associadas**, 2003. Disponível em: http://araguaia2.ufmt.br/professor/disciplina_arquivo/100/20131016140.pdf. Acesso em Junho de 2017.

FLORIDI, L. **Information: A Very Short Introduction. New York: Oxford University Press**, 2010.

FUSCO, E.; MUCHERONI, M. L.; CONEGLIAN, C. S., **Plataforma Informacional do Ecossistema Paulista de Inovação: Modelo Computacional e Semântico De Apoio à Inovação**. In: XVIII Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB, Marília, São Paulo 2017.

GLIOZZO, A.; PATWARDHAN, S.; BIRAN, O.; MCKEOWN, K. **Semantic Technologies in IBM Watson**, Proceedings of the Fourth Workshop on Teaching Natural Language Processing, Bulgaria, 2013, p. 85-92.

GRUBER, T. R. **A translation approach to portable ontology specifications**. Knowledge acquisition 5.2. 199-220. 1993.

GRUBER, T. **Ontolingua: A mechanism to support portable ontologies**, 1992.

GUARINO, N**. Formal ontology in information systems**. Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy. Vol. 46. IOS press, 1998.

HAN, J.; LE, G.; E, H.; DU, J. **Survey on NOSQL Database** (2011).

HEDLEY, J. **Jsoup: Java HTML Parser**. 2009-2016. ed. [S.l.], 2016.

HSU, J. **Why big data will have a big impact on sustainability**, 2014. Disponível em The Guardian: http://www.theguardian.com/sustainable-business/big-data-impact-sustainable-business. Acesso em: 09 de junho de 2017.

IBM. International Business Machines. **Big Data University Beta**, 2012. Disponível em: https://www.bigdatauniversity.com.br/. Acesso em: 09 de junho de 2017.

IBM Watson. **Natural Language Understand - API Reference**, 2017. Acessado em: Novembro de 2017. Disponível em: https://www.ibm.com/watson/developercloud/natural-language-understanding/api/v1/. Acesso: Novembro 2017.

IBM Watson. **Language Translate - API Reference**. Acessado em: Novembro de 2017. Disponível em: https://www.ibm.com/watson/developercloud/language-translator/api/v2/. Acesso: Novembro 2017.

MAURO, A.; GRECO, M.; GRIMALDI M. **What is Big Data? A Consensual Definition and a Review of Key Research Topics,** 2014**.**

MALIK, S. K.; RIZVI R., **Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation**, International Conference on Computational Intelligence and Communication Systems, 2011.

MANYIKA, J. et al., Big data: **The next frontier for innovation, competition, and productivity. McKinsey Global Institute**, 2011. Disponível em: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation. Acesso: Maio de 2017.

MARCONDES, C. H. **Metadados: descrição e recuperação na Web**, In: MARCONDES, C. H. et al. (Org.). Bibliotecas digitais: saberes e práticas. Salvador, BA: EDUFBA, Brasília, IBICT, 2005. p. 77-143.

MARCONDES, C. H. **“LINKED DATA” – DADOS INTERLIGADOS - E INTEROPERABILIDADE ENTRE ARQUIVOS, BIBLIOTECAS E MUSEUS NA WEB**, 2012.

MENDONÇA, Eduardo. **Extração Resiliente de Dados RDF a partir de Fontes Dinâmicas em Linguagem de Marcação**, dissertação de mestrado para o programa de mestrado da Universidade Federal do Ceará, Fortaleza, Ceará, 2003.

MORATO, A. C., MORAES, M. A. **METADADOS, DUBLIN CORE: UMA BREVE INTRODUÇÃO**. Disponível em: http://eprints.rclis.org/14424/1/Dublin_Core_-_uma_breve_introdu%C3%A7%C3%A3o.pdf. Acesso em: 09 de junho de 2017.

MOURA, A. L. T, AMORIM, D. G. **BIG DATA: O IMPACTO E SUA FUNCIONALIDADE NA SOCIEDADE TECNOLÓGICA**, 2014. Disponível em: http://revistaopara.facape.br/article/view/121/72. Acesso em 09 de junho de 2017.

NOY, N. F., E MCGUINNESS, D. L. **Ontology development 101: A guide to creating your first ontology**. 2001.

PEREIRA, F. D. **Automação do fluxo Informacional entre atores de inovação no Brasil para processos de tomada de decisão**. 2016.

PICKLER, M. E. V. **Web Semântica: ontologias como ferramentas de representação do conhecimento,** Perspect. Ciência da Informação vol.12 no.1 Belo Horizonte Jan./Apr. 2007.

RAMALHO, R. A. S. **Web Semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação**, dissertação, UNESP, Marília, 2006.

SANTAREM SEGUNDO, J. E.; VIDOTTI, S. A. B. G. **Rede de tags para recuperação da informação no contexto da Representação Iterativa**. InCID: Revista de Ciência da Informação e Documentação, v. 2, n.1, p. 86-109, 2011

SILVA, T. M. S. **Extração de informação para busca semântica na web baseada**, 2003. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/85791. Acesso: Maio de 2017.

SCHROECK, M. et al. **Analytics: The real-world use of big data**, 2012.

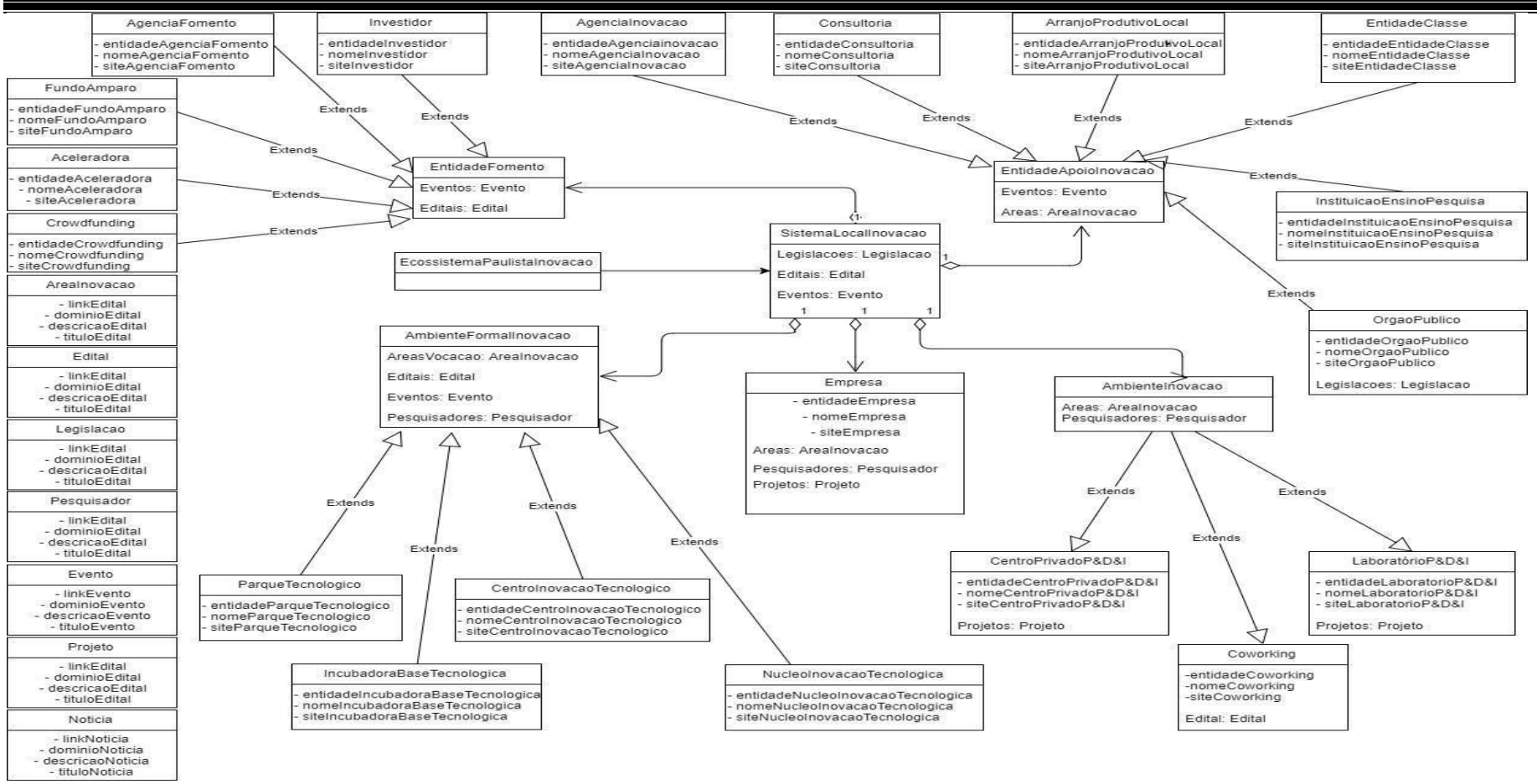
SPERONI, R. M. **Minicurso: Introdução ao RDF e SPARQL**, 2014. Acessado em: Novembro de 2017. Disponível em: http://www.inf.ufsc.br/~jose.todesco/LODBrasil/Minicurso/rdf_sparql.pdf

SOUZA, M. I. F.; VENDRUSCULO, L. G.; MELO, G. C. **Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core**. Ciência da Informação, v. 29, n. 1, p. 93-102, jan./abr. 2000.

WARD, J. S., BARKER, A. **Undefined By Data: A Survey of Big Data Definitions**, 2013.

Apêndice A*

UML ALTERADO DA INOVAONTO



Fonte: UML alterado do InovaOnto (Adaptado de Fusco & Mucheroni & Coneglian (2017))