Newton Juniano Calegari

Proposta de uma ferramenta de anotação semântica para publicação de dados estruturados na Web

São Paulo

2016

Newton Juniano Calegari

Proposta de uma ferramenta de anotação semântica para publicação de dados estruturados na Web

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Tecnologias da Inteligência e Design Digital, na área de concentração de Processos Cognitivos e Ambientes Digitais, na linha de pesquisa Design Digital e Inteligência Coletiva sob a orientação do Prof. Dr. Demi Getschko.

Pontifícia Universidade Católica de São Paulo

Tecnologias da Inteligência e Design Digital

Programa de Pós-Graduação

Orientador: Demi Getschko

São Paulo

2016

Newton Juniano Calegari

Proposta de uma ferramenta de anotação semântica para publicação de dados estruturados na Web

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Tecnologias da Inteligência e Design Digital, na área de concentração de Processos Cognitivos e Ambientes Digitais, na linha de pesquisa Design Digital e Inteligência Coletiva sob a orientação do Prof. Dr. Demi Getschko.

Prof. Dr. Demi Getschko
Orientador

Prof. Dr. Seiji Isotani
ICMC-USP

Prof. Dr. Diogo Cortiz da Silva
PUC-SP

São Paulo
2016

À minha mãe, Antônia, por todo seu amor e carinho.

Agradecimentos

A Deus, por ter colocado pessoas maravilhosas em minha vida.

À minha família e à minha namorada, Caroline, pelo constante apoio e incentivo durante meu percurso acadêmico.

Ao Prof. Dr. Demi Getschko, pelos valiosos ensinamentos e pela oportunidade concedida.

Ao Prof. Dr. Diogo Cortiz, pelas contribuições por meio das ricas conversas.

À Sra. Edna Conti, pela inestimável ajuda ao longo desta jornada.

À Pontifícia Universidade Católica de São Paulo e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão do auxílio CAPES/PROSUP, na modalidade bolsa taxa.

Começar é parte mais difícil de um trabalho.
Platão

Resumo

A proposta apresentada nesta pesquisa busca aproximar as tecnologias de Web Semântica dos usuários publicadores de conteúdo na Web, permitindo que estes contribuam com a geração de dados estruturados e metadados sobre textos e informações que venham a disponibilizar na Web. O objetivo geral deste trabalho é investigar a viabilidade técnica de desenvolvimento de uma ferramenta de anotação semântica que permita aos usuários publicadores de conteúdo contribuir para o ecossistema de Web Semântica. Com suporte de (BERNERS-LEE et al., 2001; ALESSO; SMITH, 2006; RODRÍGUEZ-ROCHA et al., 2015; GUIZZARDI, 2005; ISOTANI; BITTENCOURT, 2015) é apresentado o tópico de Web Semântica, de acordo com a pilha tecnológica que mostra o conjunto de tecnologias propostas para a sua realização. Considerando a importância de ontologias e vocabulários para a construção de aplicações de Web Semântica são apresentados, então, os tópicos fundamentais de modelagem conceitual e a linguagem de ontologias para Web. Para fornecer a base necessária para a utilização de anotações semânticas são apresentados, além da definição, os modos de uso de anotações (manual, semi-automático e automático) e as formas de integrar essas anotações com recursos disponíveis nas tecnologias da Web Semântica. O estado da arte contempla trabalhos e projetos recentes sobre o uso de Web Semântica, no contexto de publicação de conteúdo na Web. O estado da arte contempla trabalhos e projetos recentes sobre o uso de Web Semântica no contexto de publicação de conteúdo na Web. A metodologia é baseada na proposta apresentada por SANTAELLA; VIEIRA (2008), seguindo uma abordagem exploratória para a condução da pesquisa. São apresentados a proposta e os componentes de uma ferramenta de anotação semântica que utiliza vocabulários compartilhados para a geração de dados estruturados, a partir de conteúdo textual. Concluindo o trabalho, são apresentadas as possibilidades futuras, tanto da implementação da ferramenta em um cenário real, atestando sua viabilidade técnica, quanto novos trabalhos encaminhados a partir desta pesquisa.

Palavras-chaves: web semântica, dados estruturados, anotação semântica.

Abstract

The tool proposed in this research aims at bringing together the Semantic Web technologies and content publishers, this way enabling the latter to contribute to creating structured data and metadata about texts and information they may make available on the Web. The general goal is to investigate the technical feasibility of developing a semantic annotation tool that enables content publishers to contribute to the Semantic Web ecosystem. Based on (BERNERS-LEE et al., 2001; ALESSO; SMITH, 2006; RODRÍGUEZ-ROCHA et al., 2015; GUIZZARDI, 2005; ISOTANI; BITTENCOURT, 2015), the Semantic Web is presented according to its technological stack. Considering the importance of the ontologies and vocabularies used to create Semantic Web applications, the essential subjects of the conceptual modelling and the ontology language used on the Web are presented. In order to provide the necessary concepts to use semantic annotations, this dissertation presents both the way annotations are used (manual, semi-automatic, and automatic) as well as the way these annotations are integrated with resources available on the Web. The state-of-the-art chapter describes recent projects and related work on the use of Semantic Web within Web-content publishing context. The methodology adopted by this research is based on (SANTAELLA; VIEIRA, 2008; GIL, 2002), in compliance with the exploratory approach for research. This research presents the proposal and the architecture of the semantic annotation tool, which uses shared vocabulary in order to create structured data based on textual content. In conclusion, this dissertation addresses the possibilities of future work, both in terms of the implementation of the tool in a real use case as well as in new scientific research.

Key-words: semantic web, structured data, semantic annotation.

Lista de ilustrações

Lista de abreviaturas e siglas

HTTP	Hypertext Transfer Protocol
HTML	Hypertext Markup Language
IA	Inteligência Artificial
OBIE	Ontology-Based Information Extraction
OWL	Web Ontology Language
RDF	Resource Description Framework
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WWW	World Wide Web

Sumário

Introdução

A Web Semântica tem como objetivo prover a representação de dados na Web com uma semântica atribuída a eles, a fim de permitir que máquinas processem esses dados, com base na interpretação de sua informação (ALESSO; SMITH, 2006; RODRÍGUEZ-ROCHA et al., 2015).

O tema foi apresentado por BERNERS-LEE et al. (2001) em um artigo publicado na revista *Scientific American*. Este artigo seminal deu origem a diversas pesquisas que têm sido realizadas em direção à criação de padrões e tecnologias para ampliar os modos de uso da Web - deixando de ser apenas um repositório de dados para também comportar bases de conhecimento, em uma escala global.

Os padrões e tecnologias que vêm sendo criados para suportar a implementação da Web Semântica permitem que novos tipos de aplicações sejam desenvolvidas. Essas tecnologias fornecem a estrutura necessária para que aplicações sejam capazes de processar e interpretar os dados disponíveis na Web, considerando o significado atribuído a eles. De acordo com KARGER (2014), "a Web Semântica tem o potencial de revolucionar a maneira com que usuários finais obtém, comunicam e gerenciam informações" (nossa tradução), porém ainda não atingiu esse potencial, possivelmente devido ao fato de que não sejam muitas as aplicações de Web Semântica disponíveis para todas as escalas de usuários na Web.

SLIMANI (2013) considera que a disponibilização de metadados e de dados estruturados sobre os recursos na Web é um fator importante para a realização da Web Semântica, mas o sucesso dessa abordagem depende diretamente de um grande número de usuários, gerando e explorando dados estruturados na Web. Para que mais usuários contribuam para este processo, parece ser necessário tornar simples, em aspectos de usabilidade, as ferramentas de Web Semântica (Ibid.).

Para a condução desta pesquisa, foram adotados o método exploratório descrito por (GIL, 2002) e a metodologia apresentada por (SANTAELLA; VIEIRA, 2008). As abordagens indicadas pelos autores guiam para um levantamento bibliográfico preliminar, aproximando o pesquisador do problema pesquisado, com o objetivo de familiarizá-lo com o tema e de facilitar-lhe a formulação de hipóteses. O método exploratório apresentado por GIL (2002, p.41) consiste em etapas de levantamento bibliográfico, entrevistas e contato com pessoas que tiveram proximidade com o problema apontado na pesquisa, além de análises de exemplos que estimulem a compreensão do objeto, tarefa esta que é contemplada no estado da arte desta pesquisa, o qual permite uma maior contextualização com o tema.

Os tópicos fundamentais que formam a base necessária da Web Semântica são apresentados por (BERNERS-LEE et al., 2001), (GUIZZARDI, 2005), (GRUBER, 1993), (MANOLA; MILLER; MCBRIDE, 2004), (ISOTANI; BITTENCOURT, 2015).

Os conteúdos relacionados com anotação semântica são destacados por (SLIMANI, 2013), (UREN et al., 2006), (ANDREWS; ZAIHRAYEU; PANE, 2011), (RODRÍGUEZ-ROCHA et al., 2015), entre outros. Esta pesquisa se mostra pertinente e relevante entre as propostas apresentadas pelos autores que tratam do tema de anotação semântica.

A proposta aqui apresentada se faz oportuna e apropriada, pois se baseia no tópico de anotações semânticas, com o objetivo de possibilitar que um número maior de usuários finais passem a usufruir das tecnologias de Web Semântica, as quais,segundo (KARGER, 2014), possuem potencial para lidar, entre outros, com os problemas de organização de informação enfrentado pelos usuários; porém, pouco tem sido explorado nesse caminho.

O problema que tratamos traz a questão de como possibilitar que os usuários publicadores de conteúdo passem a produzir mais conteúdos semânticos na Web, utilizando vocabulários e ontologias para descrever os dados a serem publicados; partindo da hipótese de que este tipo de usuário pode contribuir com a geração de metadados e dados estruturados por meio do uso de uma tecnologia que possa ser integrada à sua rotina de escrita e publicação de conteúdo na Web, esta pesquisa tem como objetivo estudar e investigar a viabilidade técnica de desenvolvimento de uma ferramenta de anotação semântica que permita aos usuários publicadores de conteúdo contribuir para o ecossistema de Web Semântica.

Para permitir que mais usuários, incluindo aqueles publicadores de conteúdo, participem do ecossistema de Web Semântica é proposta uma aplicação, com aspectos de usabilidade que tornam simples o seu emprego, viabilizando o seu aproveitamento por usuários que escrevem e publicam conteúdo textual na Web. Esta ferramenta permite realizar anotações semânticas em textos, utilizando vocabulários compartilhados para atribuir significado aos termos neles selecionados. À medida em que os termos são anotados semanticamente, a aplicação gera triplas RDF[1] e as armazena em um banco de dados projetado para suportar a estrutura de triplas. A aplicação busca abstrair os conceitos mais complexos de Web Semântica, a fim de diminuir os pré-requisitos técnicos para sua utilização.

Os tópicos que fazem parte do contexto do trabalho são apresentados a partir da explicação das tecnologias fundamentais para a Web.

O conjunto de tecnologias da Web Semântica é explicado de acordo com a ilustração que classifica em camadas os assuntos pertinentes à área. Ontologias e vocabulários e os tipos de anotações semânticas são apresentados nas duas últimas seções do capítulo

[1] *Resource Description Framework*

de fundamentação teórica.

O resultado esperado desta pesquisa é a apresentação de um modelo e protótipo de uma aplicação para realização de anotações semânticas, em conteúdo textual. A intenção, com esta proposta, é de aproximar as tecnologias de Web Semântica à determinados grupos de usuários finais, como jornalistas, editores de conteúdo e redatores, pois, como afirma KARGER (2014, p.64), pouco tem sido realizado nesse sentido, o de trazer o potencial da Web Semântica aos usuários finais.

> *The Semantic Web's potential to deliver tools that help end users capture, communicate, and manage information has yet to be fulfilled, and far too little research is going into doing so.*
>
> O potencial da Web Semântica para entregar ferramentas que ajudem os usuários finais a capturar, comunicar e gerenciar informações ainda deve ser realizado e muito pouco tem sido pesquisado nesse sentido. (nossa tradução)

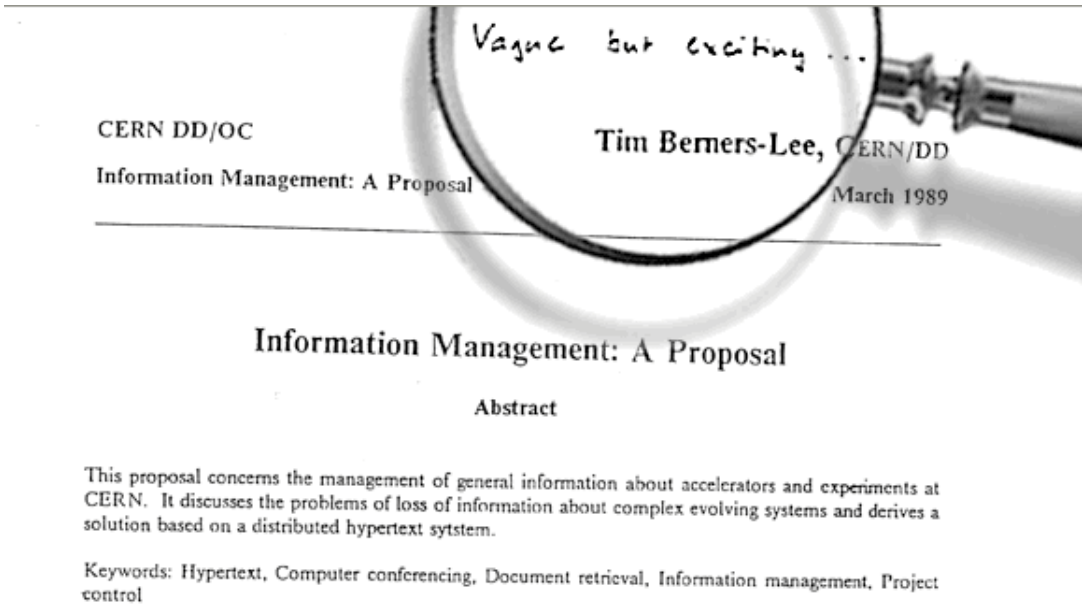
Os detalhes do protótipo são descritos no terceiro capítulo da dissertação , no qual são apresentadas a arquitetura da aplicação, as tecnologias sugeridas para o seu desenvolvimento e o possível cenário para implementação.

1 Fundamentação teórica

Neste capítulo são apresentadas as tecnologias fundamentais para criação da *World Wide Web*, a definição de Web Semântica e os elementos que compõem o conjunto de tecnologias relacionadas, além dos conceitos sobre ontologias e vocabulários, importantes para a Web Semântica e o embasamento teórico necessário para a aplicação de anotações semânticas.

1.1 Os fundamentos da *World Wide Web*

O projeto que deu origem à Web nos moldes utilizados na atualidade, foi desenvolvido por Tim Berners-Lee no CERN[1]. Um problema existente no CERN ao fim da década de 80 dizia respeito ao gerenciamento das informações geradas por diversos cientistas que conduziam suas pesquisas no laboratório (BERNERS-LEE, 1989)e foi a partir deste problema que Berners-Lee propôs, em março de 1989, um modelo para um sistema global de hipertexto. O desenvolvimento do projeto foi iniciado após a aprovação por Mike Sendall, então chefe de Berners-Lee, no centro de pesquisas, com a utilização da seguinte frase, ao se referir à proposta apresentada: *"Vague, but exciting"*.


Vague but exciting ...

CERN DD/OC

Information Management: A Proposal

Tim Berners-Lee, CERN/DD

March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Figura 1: Imagem da proposta de Tim Berners-Lee com a anotação "*Vague, but exciting*". Fonte (CERN, 2008)

[1] "Organização Europeia para a Pesquisa Nuclear", nome derivado do acrônimo em francês para *Conseil Européen pour la Recherche Nucléaire*.

A solução apresentada para lidar com o acesso à informação no CERN faz menção ao conceito de hipertextos, apresentado por NELSON (1967). O termo hipertexto, contudo, foi utilizado para transmitir duas ideias diferentes: a primeira diz a respeito à conexão entre informações legíveis e passíveis de interpretação por humanos[2], de maneira irrestrita ou não-linear, enquanto a segunda não se refere apenas a textos, mas também a documentos multimídia, que incluem gráficos, áudios e vídeos que, neste caso, são utilizados sob o termo hipermídia (do inglês *Hypermedia*) (NELSON, 1967 apud BERNERS-LEE, 1989).

A Web é hoje tão popular e ubíqua que, não raro, acaba sendo confundida com a própria Internet (SOUZA; ALVARENGA, 2004, p.132). A Internet é uma infraestrutura de redes, servidores e sistemas autônomos ou ainda, de acordo com a definição de KUROSE; ROSS (2010, p.2) "a Internet pode ser descrita em termos de uma infraestrutura de redes que fornece serviços para aplicações distribuídas", e a Web se destaca como sendo uma destas aplicações distribuídas, funcionando sobre infraestrutura de redes da Internet.

Na década de 90, a Internet já havia se tornado o principal meio de conectar computadores entre diferentes redes acadêmicas e comerciais, porém não havia atingido a escala de uma rede de dimensões globais, com milhões de nós, como atualmente. Possivelmente, a criação da WWW tenha favorecido o crescimento da Internet como presenciamos atualmente, pois introduziu um novo modo para que pessoas se conectassem e trocassem informações na rede.

A criação da WWW como um sistema global de hipertexto foi possível devido ao desenvolvimento de três pilares fundamentais para a aplicação: o protocolo HTTP, a linguagem HTML e a sintaxe genérica para identificação de recursos, URI.

O HTTP[3] é um protocolo no nível de aplicação utilizado por sistemas distribuídos, colaborativos e sistemas de informações hipermídias (FIELDING et al., 1999), para a troca de documentos de hipertexto e hipermídia, através da Internet. Embora tivesse sido especificado como um padrão na RFC 1945 (1996), o protocolo HTTP já vinha sendo utilizado para transferência de documentos de hipertexto na Web, desde o início dos anos 90.

O protocolo HTTP define como determinado computador cliente solicita um documento para o servidor e como o servidor transfere esse documento para o computador cliente; desse modo, são duas aplicações implementando o mesmo protocolo: uma aplicação no computador cliente (no navegador web, por exemplo) e outra no servidor (em um servidor web, *Apache*, por exemplo). A comunicação entre eles ocorre, geralmente, pelas portas 80 e 443. A porta 443 permite a comunicação criptografada pelo protocolo HTTP,

[2] Optou-se pela frase "passível de interpretação por humanos" para se referir ao termo termo *human-readable*.

[3] Hypertext Transfer Protocol

com uma camada de segurança, por meio de HTTPS. (KUROSE; ROSS, 2010, p.72)

A linguagem HTML, baseada na linguagem de marcação SGML (*Standard Generalized Markup Language*), padrão internacional *ISO 8879:1986*, é utilizada para definir a estrutura de documentos de hipertexto - páginas web - e é interpretada pelos navegadores. As discussões a respeito de padronização e o surgimento de novos navegadores capazes de processar documentos escritos em HTML aconteceram a partir da criação da lista de discussões *www-talk*[4], em 1991. A linguagem HTML se tornou propícia para que documentos fossem referenciados de maneira não-linear, na Web, ao incorporar em seu conjunto de *tags* os links de hipertexto, utilizando o elemento âncora (`<a>`) com o atributo `href`.

> *Certainly the simplicity of HTML, and the use of the anchor element (`<a>`) for creating hypertext links, was what made Tim's invention so useful.*
>
> Certamente, a simplicidade do HTML e o uso do elemento âncora (`<a>`) para a criação de links de hipertexto, foram os responsáveis pela invenção de Tim tão útil. (nossa tradução)

RAGGETT et al. (1998) afirma que a simplicidade da linguagem HTML e a utilização do elemento `<a>` para links de hipertexto foram importantes para mostrar a utilidade do projeto proposto por Berners-Lee.

Com a criação da *World Wide Web*, Berners-Lee continuara "uma tradição de mais de 50 anos na busca de soluções para associar fontes de informação por meio da computação interativa, iniciadas por Bush, Engelbart, Ted Nelson e Atkinson" (GARCIA, 2011, p.29), e ainda, como apontado por (CASTELLS, 2003, p.18), o projeto da WWW progrediu, pois teve a vantagem decisiva de que a Internet já existia na época, e pôde, desse modo, apoiar-se em sua estrutura e poder computacional descentralizado.

A dimensão que a Web atingiu, tanto no número de recursos quanto no número de pessoas interagindo, fazem dela um excelente instrumento para interação social. Tim Berners-Lee (2000, p.123), afirma que a Web foi criada não como um "brinquedo tecnológico", mas sim para um efeito social - o de ajudar para que pessoas trabalhem juntas.

> *The Web is more a social creation than a technical one. (BERNERS-LEE; FISCHETTI, 2000, p.123)*
>
> A Web é mais uma criação social do que uma invenção técnica. (nossa adaptação)

Com essa semente plantada - a da criação de uma ferramenta com propósito social - a Web surgiu e evoluiu a ponto de oferecer muitos outros frutos, como as diferentes

[4] Lista de discussão *www-talk*: <https://lists.w3.org/Archives/Public/www-talk/>

aplicações baseadas na Web, além das diversas ramificações tecnológicas que surgem a fim de melhorá-la como um todo. Entre tais desdobramentos de aspectos tecnológicos, surgiu a abordagem da Web Semântica, encarada como um complemento para a Web de Documentos, apresentada nesta seção.

1.2 Web Semântica

O artigo seminal *"The Semantic Web"* publicado pela revista *Scientific American* em 2001 e escrito por Tim Berners-Lee, James Hendler e Ora Lassila marcou o início das pesquisas e contribuições para a evolução da Web em direção a um ambiente no qual agentes de software interpretam dados nela disponíveis, com base nos seus significados, passando a auxiliar usuários na realização de diversas tarefas no ambiente digital.

A Web Semântica não é uma Web separada da "atual" - a Web de documentos - mas sim uma extensão que busca prover aos sistemas e agentes de software a capacidade de processar, compartilhar, reusar e entender os termos descritos por dados definidos de maneira precisa (BERNERS-LEE et al., 2001; ISOTANI; BITTENCOURT, 2015).

Uma definição recente para complementar o entendimento da Web Semântica é apresentada por RODRÍGUEZ-ROCHA et al. (2015, p.33):

> *The Semantic Web aims to describe the meaning of information published on the Web to enable retrieval based on an accurate understanding of that information's semantics.*
>
> A Web Semântica tem por objetivo descrever o significado da informação publicada na Web com a finalidade de habilitar a realização de consultas baseadas no entendimento preciso da semântica daquela informação. (nossa tradução)

A abordagem da Web Semântica é utilizada para tornar os conteúdos relevantes de diferentes páginas passíveis de interpretação por máquinas[5], permitindo que algoritmos e agentes de software executem tarefas de modo automatizado, consultando dados com base em sua semântica, para atender a demanda dos usuários (BERNERS-LEE et al., 2001).

A Web de documentos é caracterizada principalmente pela capacidade de interligação de documentos por meio dos *links* de hipertexto, enquanto a Web Semântica, ou Web de dados, busca atribuir semântica aos dados e recursos disponibilizados na Web, interligando-os com o uso de tecnologias como RDF.

O termo *Web of Data* refere-se a um espaço global contendo bilhões de fatos, no qual a Web se faz o ambiente interoperável para esse gigantesco banco de dados.

[5] "passível de interpretação por máquina" nossa adaptação do termo em inglês *machine-readable*

Embora existam diferentes interpretações para a visão de uma Web Semântica, como afirma (BIZER; HEATH; BERNERS-LEE, 2009), o objetivo de construir uma Web com cada vez mais dados *machine-readable* permanece constante.

> *The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web – a web of data that can be processed directly or indirectly by machines.*
> *(BERNERS-LEE; FISCHETTI, 2000 apud BIZER; HEATH; BERNERS-LEE, 2009)*

> O primeiro passo é publicar dados na Web, de modo que máquinas possam entendê-los naturalmente ou convertê-los para um formato adequado. Isso cria o que chamamos Web Semântica, ou seja, uma web de dados que pode ser processada direta ou indiretamente por máquinas. (nossa tradução)

Sob a denominação de *Linked Data* há um conjunto de princípios, boas práticas e recomendações para publicação e conexão de dados estruturados na Web. Essa abordagem contribui para que seja atingido o objetivo da construção de uma Web de Dados, na qual dados estruturados, com semântica explícita e em formato *machine-readable* estejam disponíveis.

Um conjunto de tecnologias e padrões para implementação de tecnologias de Web Semântica foi proposto e organizado conforme a estrutura da chamada "pilha de tecnologias da Web Semântica", mostrada na figura 2, também apresentada como *Semantic Web Stack* ou *Semantic Web Layercake*. Além dos padrões e tecnologias existentes, já utilizados para a formação da Web de documentos, outros foram propostos a fim de contribuir para a realização da Web de dados.

A Web de dados, por ser uma extensão da Web de documentos, mantém o mesmo conjunto de padrões fundamentais (HTTP, HTML e URI) para servir como base para os outros padrões representados na "pilha tecnológica". Localizado na base da estrutura mostrada na figura 2 o padrão *Unicode* provê um conjunto de caracteres universais que habilitam as aplicações a serem disponibilizadas em quaisquer idiomas, independentemente do conjunto de caracteres utilizados pela língua.

A identificação global dos recursos acontece por meio de URI (*Uniform Resource Identifiers*). Para que seja possível o compartilhamento de informações entre os recursos, os mesmos devem possuir identificadores únicos. Quando se faz necessária a internacionalização de caracteres por meio de Unicode para esses identificadores, são atribuídos os IRI (*Internationalized Resource Identifiers*). Por exemplo, quando há informação sobre a localização, em termos de protocolo de determinado recurso identificado, o conjunto de identificadores utilizado é chamado URL (*Uniform Resource Locator*). Uma tentativa de clarificar a distinção entre os três conceitos é apresentada a seguir:

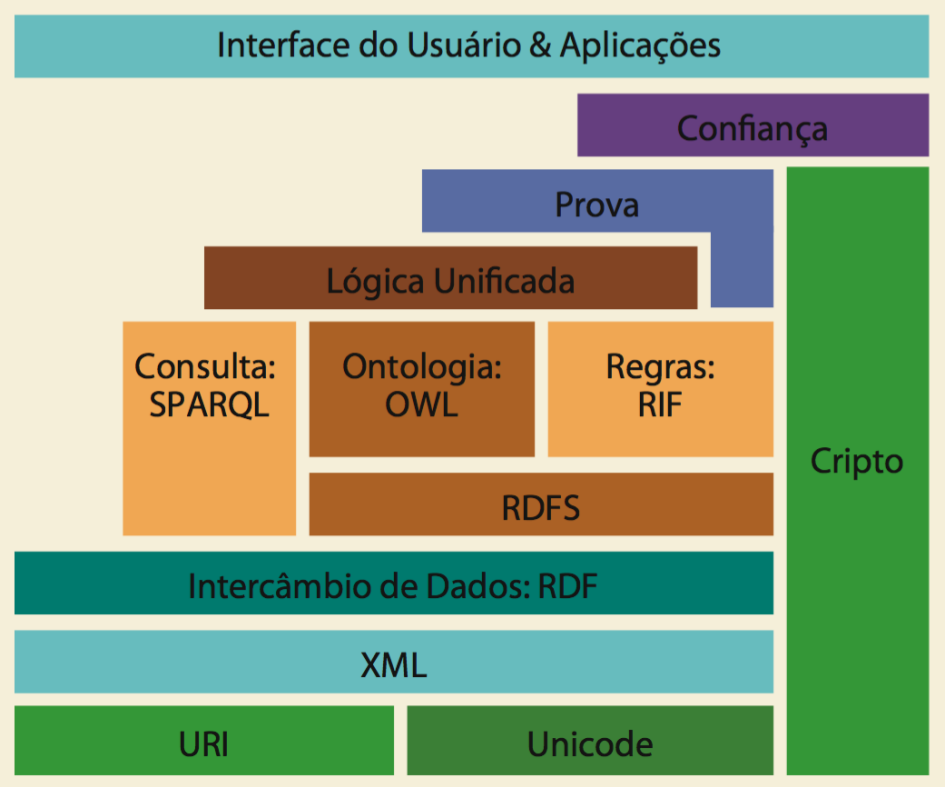


Figura 2: Conjunto de tecnologias da Web Semântica. Fonte: ISOTANI; BITTENCOURT (2015)

URI É uma sequência de caracteres que identifica um recurso físico ou abstrato. Teve sua sintaxe genérica definida na RFC 2396 (1998).

IRI A RFC 3987 (2005) propõe a padronização de uma extensão para URI, utilizando o conjunto de caracteres universais (Unicode/ISO 10646) a fim de permitir a internacionalização dos identificadores de recursos.

URL É um subconjunto do conjunto de URI. As URL fornecem a localização dos recursos descrevendo seu mecanismo de acesso primário. Portanto, quando recursos identificados na Web possuírem informação referente ao protocolo de acesso, seus identificadores serão definidos como URL.

A próxima camada apresenta a linguagem XML (*Extensible Markup Language*) como a linguagem de marcação para criação de documentos compostos de dados estruturados. XML permite a utilização de diferentes *namespaces* para tornar explícito o contexto das diferentes *tags*, em seus documentos, tornando possível, por meio de *XML Namespaces*, a utilização de marcações provenientes de diversas fontes. Há outras linguagens e tecnologias com o mesmo propósito de formalizar a estrutura de documentos para Web Semântica, como a especificação JSON-LD (SPORNY; KELLOGG; LANTHALER, 2013).

Acima da camada propondo a tecnologia necessária para a criação de documentos com dados estruturados, há a camada responsável por propor o ferramental necessário para o intercâmbio de dados. Para isso, considera-se o modelo RDF, projetado para descrever recursos na Web. Este modelo RDF utiliza o padrão de representação em triplas baseadas no conjunto *Sujeito, Predicado, Objeto*. A recomendação do *Resource Description Framework*, pelo W3C (MANOLA; MILLER; MCBRIDE, 2004) define RDF como uma linguagem para representação de informações sobre recursos na Web.

Documentos no modelo RDF podem ser serializados em diferentes formatos, como RDF/XML, Turtle, JSON-LD, entre outros, mas se mantém comum nesses diferentes formatos de serialização a estrutura de tripla, com os elementos *sujeito, predicado e objeto*, como representado no grafo da figura 3:

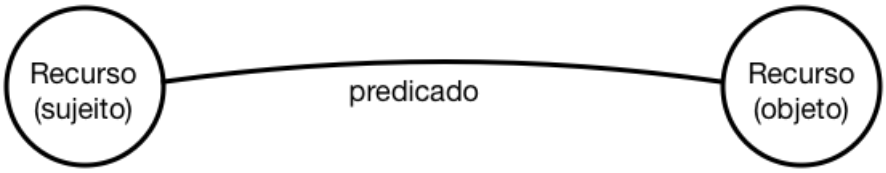


Figura 3: Tripla com *sujeito*, *predicado* e *objeto* representada em um grafo

A aresta representa um predicado ou uma propriedade, enquanto os nós representam os recursos.

A figura 4 exibe um grafo RDF com informações sobre determinada pessoa, no qual os predicados das triplas são identificados por URI.

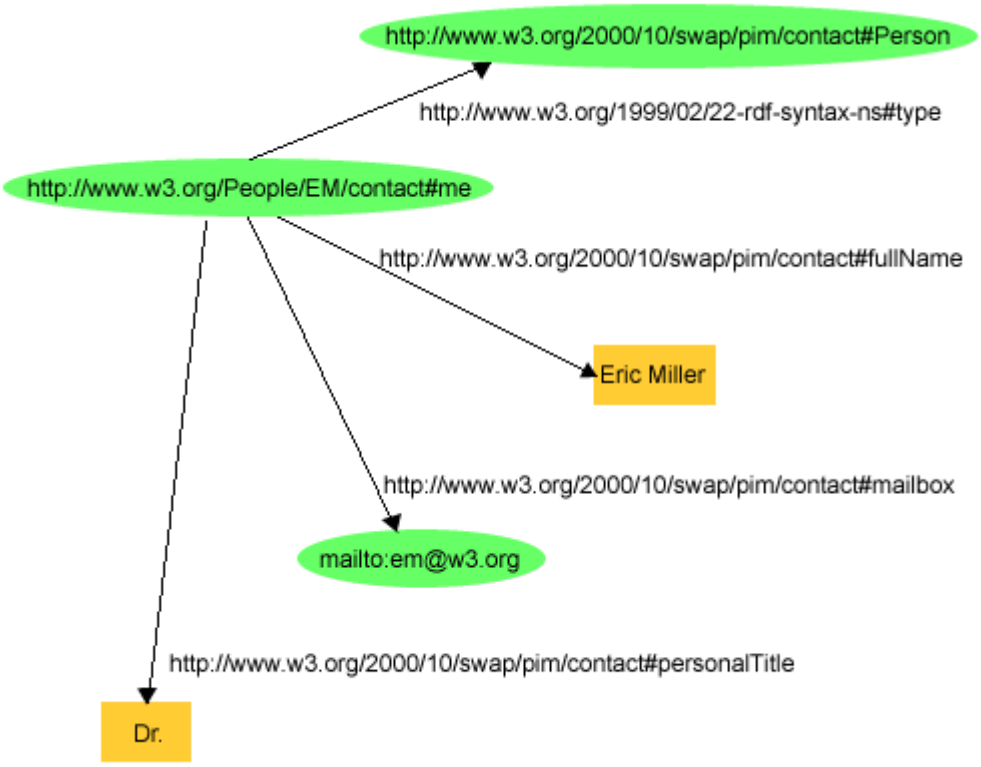


Figura 4: Grafo RDF - Adaptado de (MANOLA; MILLER; MCBRIDE, 2004)

Para realização de consultas aos dados modelados no padrão RDF, mais precisamente em grafos RDF, é utilizada a linguagem SPARQL.

As consultas com uso da linguagem SPARQL são realizadas em uma arquitetura cliente servidor, transferindo o conteúdo da *query* do cliente para um servidor chamado *SPARQL end-point*, que geralmente possui um banco de dados para armazenamento de triplas RDF.

O acrônimo SPARQL significa *SPARQL Protocol and RDF Query Language*. A especificação mais recente da linguagem trata da versão 1.1 e foi publicada como recomendação pelo W3C em Março de 2013.

SPARQL pode ser utilizada para realização de consultas a diversas fontes de dados, nas quais estes dados estejam nativamente no padrão RDF ou possam ser convertidos para RDF, por meio de alguma outra aplicação *middleware*. A especificação 1.1 da linguagem SPARQL provê suporte a funções de agregação, *subqueries*, negação, filtros de valores, entre outras funcionalidades (HARRIS; SEABORNE, 2013). Uma consulta SPARQL é realizada sobre um conjunto de dados em RDF e o resultado desta consulta pode ser outro grafo RDF.

Uma consulta simples em SPARQL consiste de duas partes básicas:

```
SELECT ...
WHERE { ... }
```

A cláusula *SELECT* identifica as variáveis, sejam elas sujeito, predicado ou objeto, que deverão ser retornadas na consulta.

E na cláusula *WHERE* são definidos os padrões das triplas a serem consultados nos grafos RDF.

Além dessas duas cláusulas, uma consulta SPARQL simples pode ter outros dois blocos úteis em diferentes casos: o bloco *PREFFIX*, no qual podem ser declaradas as abreviações dos URI dos esquemas (ontologias e vocabulários) utilizados na consulta, e a cláusula *FROM*, na qual podem ser identificadas as fontes de dados a serem utilizadas na consulta, como o URI de um arquivo no formato *Turtle*, contendo um grafo RDF.

No exemplo abaixo, é realizada uma consulta com a linguagem SPARQL na base de dados da DBPedia[6].

O objetivo desta consulta é retornar todas as obras do pintor holandês Vincent van Gogh. O repositório de dados da DBPedia é formada com base nos dados disponíveis na Wikipedia.

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbr: <http://dbpedia.org/resource/>

SELECT ?obra
WHERE {
        ?obra dbp:artist dbr:Vincent_van_Gogh
}
```

A figura 5 mostra um grafo representando a tripla RDF definida na consulta às obras de Van Gogh. Nesta tripla os elementos são: `?obra` (sujeito), `dbp:artist` (predicado), `dbr:Vincent_van_Gogh` (objeto).

[6] <http://dbpedia.org>

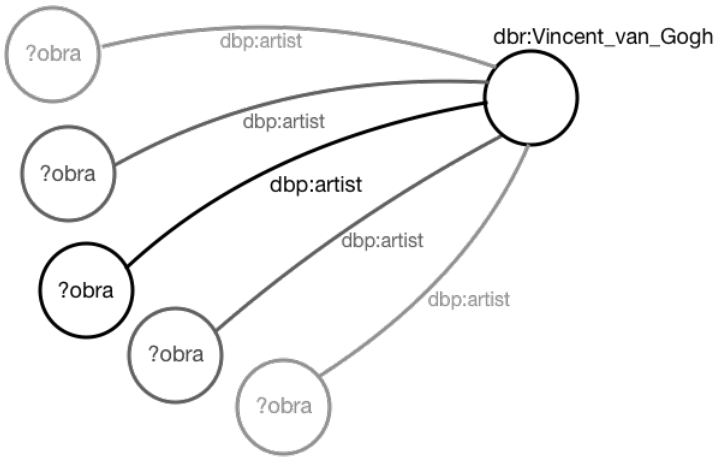


Figura 5: Representação de um grafo para uma consulta SPARQL

A consulta foi realizada na interface que permite consultas ao banco de dados RDF *Virtuoso*, por meio de um *SPARQL end-point*[7], disponibilizados pela DBPedia.

O resultado da consulta (figura 6) apresenta os URI dos recursos recursos encontrados, ou seja, os URI das obras de Vincent van Gogh.

O retorno de uma consulta *SPARQL* pode também ser exibido em outros formatos, como *CSV*, *JSON*, *RDF/XML*, *Turtle*, entre outros.

Dois aspectos das recomendações e especificações básicas que formam a base fundamental para Web Semântica foram apresentados: são eles o de modelagem de dados, com RDF e o de consulta aos dados, com a utilização de *SPARQL*. Na próxima seção são apresentados os conceitos de ontologias e vocabulários, que contemplam o aspecto de descrição dos dados estruturados.

[7] O *SPARQL end-point* da DBPedia está disponível em <http://dbpedia.org/sparql>, acesso em 19 de Dezembro de 2015.

obra
http://dbpedia.org/resource/The_Red_Vineyard
http://dbpedia.org/resource/Falling_Autumn_Leaves
http://dbpedia.org/resource/Boats_du_Rh%C3%B4ne
http://dbpedia.org/resource/A_Lane_in_the_Public_Garden_at_Arles
http://dbpedia.org/resource/Plain_near_Auvers
http://dbpedia.org/resource/Green_Wheat_Field_with_Cypress
http://dbpedia.org/resource/Road_with_Cypress_and_Star
http://dbpedia.org/resource/Butterflies_(Van_Gogh_series)
http://dbpedia.org/resource/Langlois_Bridge_at_Arles
http://dbpedia.org/resource/Arles:_View_from_the_Wheat_Fields
http://dbpedia.org/resource/Bulb_Fields
http://dbpedia.org/resource/A_Woman_Walking_in_a_Garden
http://dbpedia.org/resource/Montmartre_(Van_Gogh_series)
http://dbpedia.org/resource/Japonaiserie_(Van_Gogh)
http://dbpedia.org/resource/Portrait_of_Dr._Gachet
http://dbpedia.org/resource/Portrait_of_P%C3%A8re_Tanguy
http://dbpedia.org/resource/Peasant_Character_Studies_(Van_Gogh_series)
http://dbpedia.org/resource/Asni%C3%A8res_(Van_Gogh_series)
http://dbpedia.org/resource/A_Meadow_in_the_Mountains:_Le_Mas_de_Saint-Paul
http://dbpedia.org/resource/A_Wind-Beaten_Tree
http://dbpedia.org/resource/Torso_of_Venus_and_a_Landscape
http://dbpedia.org/resource/Lying_Cow

Figura 6: Resultado da consulta *SPARQL* na DBPedia.org exibidos em HTML

1.3 Ontologias e vocabulários

A palavra ontologia, formada pelos termos de origem grega "ontos", que significa "ser" e "logia", que significa "estudo", é definida na Filosofia como o estudo do ser e da realidade ou, de acordo com (PEIRCE; HARTSHORNE; WEISS, 1935 apud GUIZZARDI, 2005), ontologia trata do "estudo dos atributos mais gerais da realidade e dos objetos reais". Diferentemente de outras disciplinas científicas mais específicas, como biologia, física e química, que tratam das entidades em seus respectivos domínios, a área de ontologias lida com as possíveis relações transdisciplinares entre conceitos pertencentes a diferentes domínios da ciência, além de entidades reconhecidas por senso comum (GUIZZARDI, 2005).

Entretanto, para efeito deste trabalho e para o estudo das ontologias que formam o conjunto de elementos da Web Semântica, recorreremos às definições e usos de ontologias em Ciência da Computação. Nesta área os estudos ontológicos estão sob o olhar de uma disciplina chamada "Ontologias Aplicadas", do inglês *Applied Ontology* (MASOLO et al., 2003 apud GUIZZARDI, 2005), que utiliza teorias formais acerca de ontologias que podem ser desenvolvidas e aplicadas nos campos das ciências da computação e informação.

Gruber (1993) define ontologia como uma formalização explícita de um entendimento compartilhado de uma conceitualização, mas essa definição pode ser genérica e acabar não atendendo as especificidades de cada área da computação, nas quais ontologias são aplicadas. As áreas de banco de dados e sistemas de informação, engenharia de software, e inteligência artificial são as principais responsáveis por criar a demanda da aplicação de ontologias em ciência da computação (SMITH; WELTY, 2001 apud GUIZZARDI, 2005).

No contexto da Web Semântica, ontologias são utilizadas de modo a intermediar o entendimento humano sobre determinados domínios, mapeando-o em símbolos para que sejam processados pela máquina. Neste sentido, ontologias, como especificações formais, podem ser usadas para especificar semântica de domínios em símbolos sintáticos, para serem apresentados em recursos na Web (GUIZZARDI, 2005, p.71).

Ontologias estão no centro da arquitetura de Web Semântica proposta por (SHADBOLT; HALL; BERNERS-LEE, 2006) pois, segundo (ISOTANI; BITTENCOURT, 2015), "oferecem o apoio necessário para resolver alguns dos problemas que cercam a construção de tecnologias que utilizam bases de dados com representação formal (ou bases de conhecimento)". Tanto as ontologias quanto os vocabulários auxiliam na descrição semântica de dados e conjuntos de dados disponibilizados na Web.

Na Web, há uma diversidade de vocabulários que apresentam um conjunto de propriedades para descrever determinados dados; porém, as descrições são limitadas aos aspectos terminológicos. Utilizam-se as ontologias para garantir mais expressividade para

tais descrições, de modo que é possível, numa ontologia, definir restrições para os valores das propriedades, detalhar os relacionamentos dos termos e das classes, como classes disjuntas, e também definir a hierarquia das classes (MCCUINNESS, 2003). Portanto, vocabulários tratam apenas das relações terminológicas, enquanto ontologias permitem atingir certo grau de expressividade, lidando com as relações fenomenológicas entre os conceitos representados.

A linguagem OWL (do inglês, *Web Ontology Language*) é uma linguagem para representação de conhecimento de modo lógico que possui sintaxe baseada em RDF/XML e é utilizada para definição e instanciação de ontologias Web.

O conhecimento disponibilizado por meio de um arquivo escrito com OWL pode passar por um processo de inferência, para que sejam descobertos possíveis novos fatos sobre aquele domínio de conhecimento. A linguagem fornece certo grau de expressividade para representação formal de ontologias.

Como apontado por (SANTOS; CARVALHO, 2007, p.8), uma ontologia OWL pode contemplar relações de taxonomia entre classes, propriedades dos tipos de dados e descrições dos atributos de elementos das classes, propriedades do objeto e descrições das relações entre elementos das classes, instâncias das classes e instâncias das propriedades, além de restrições sobre tais relações.

Descrições equivocadas sobre a definição e aplicação de OWL existem, como afirmam ISOTANI; BITTENCOURT (2015, p.108).). Tais equívocos ocorrem devido à complexidade inerente ao termo "ontologias" e pela expressividade da linguagem OWL. Afim de facilitar o entendimento sobre OWL os autores citam HITZLER et al. (2012), destacando três pontos que não são características da linguagem:

Não é uma linguagem de programação: OWL é uma linguagem declarativa, utilizada para descrever determinado universo do discurso, de forma lógica. Quando se descreve conhecimento por meio de ontologias torna-se possível inferir novas informações sobre o conhecimento nelas representado. Esse processo de inferência é realizado por mecanismos conhecidos como *reasoners*. Detalhes deste processo não fazem parte do escopo da OWL.

Não é uma linguagem de esquema para conformidade sintática: não cabe à OWL definir como determinado documento deve ser sintaticamente estruturado. Isso pode ser realizado por *XML Namespaces*, por exemplo.

Não é um banco de dados: a semântica utilizada em OWL é diferente da utilizada em bancos de dados. Os bancos de dados são "mundos fechados" (do inglês *Closed-World Assumptions*), enquanto as ontologias são mundo aberto (do inglês *Open-World Assumptions*). Quando determinado fato não está presente em um banco de dados (mundo fechado), ele é considerado falso. Em OWL (mundo aberto), quando determinado

fato não está presente, ele é considerado desconhecido, porque pode vir a ser verdadeiro. Nas declarações abaixo os autores exemplificam as duas situações:

Declaração: <Seiji> <é cidadão> <Brasil>
Pergunta: <Ig> <é cidadão> <Brasil>?
Resposta (mundo fechado): Não.
Resposta (mundo aberto): Não sei.

Devem-se, também, considerar os seguintes aspectos básicos de modelagem de conhecimento em ao utilizar a linguagem OWL para representação (HITZLER et al., 2012):

Axiomas: são os fatos representados por uma ontologia escrita em OWL.
Entidades: elementos utilizados para referenciar objetos reais.
Expressões: combinações de entidades utilizadas para formar descrições mais complexas, a partir de expressões criadas pelos termos primitivos.

Os vocabulários e ontologias presentes na Web são disponibilizados em diferentes formatos, variando principalmente entre *Turtle*, *RDF/XML* e *OWL*. É apontada como uma boa prática a publicação da documentação e dos metadados relacionados à ontologia em um formato compreensível por humanos (LÓSCIO; BURLE; CALEGARI, 2015) e não somente a publicação do arquivo *machine-readable* da ontologia.

1.4 Anotações semânticas

Uma característica do conteúdo de Web Semântica é a disponibilização dos dados de um modo que seja passível de interpretação por máquina. Um dos meios de elevar a qualidade de um dado disponível na Web de *human-readable* para *machine-readable* é o processo chamado anotação semântica, o qual utiliza ferramentas como RDF para permitir essa mudança na característica dos dados.

O processo de anotação semântica consiste da geração de metadados para documentos, trechos de texto ou conceitos, por meio da criação de rótulos para estes, com o intuito de permitir recursos, como busca avançada, baseada em conceitos, inferências sobre conteúdos e visualização de informações baseadas em ontologias. Portanto, a anotação de um texto pode ser semântica quando são adicionadas aos textos de um documento informações sobre o seu significado ou sobre o significado de elementos que o compõe. (SÁNCHEZ; ISERN; MILLAN, 2011 apud RODRÍGUEZ-ROCHA et al., 2015, p.34)

Um exemplo de anotação semântica é descrito por UREN et al. (2006, p.15):

> *Semantic annotation formally identifies concepts and relations between concepts in documents, and is intended primarily for use by machines. For example, a semantic annotation might relate "Paris" in a text to an ontology which both identifies it as the abstract concept "City" and links it to the instance "France" of the abstract concept "Country", thus removing any ambiguity about which "Paris" it refers to.*

> Anotação semântica identifica formalmente conceitos e relacionamentos entre conceitos em documentos e é destinada, principalmente, a ser processada por máquinas. Por exemplo, uma anotação semântica pode relacionar "Paris" em um texto, a uma ontologia que identifica o conceito abstrato "Cidade" ao mesmo tempo que o conecta à instância "França", do conceito abstrato "País", eliminando, assim, qualquer ambiguidade ao que o termo "Paris" se refere.

Os modos de utilização de anotações semânticas são apresentados por SLIMANI (2013) conforme três categorias: manual, semi-automática e automática.

Anotação semântica manual é o processo utilizado para transformar recursos sintáticos, como texto puro (*plain-text*), em estruturas complexas, que possam conter mais informações e conhecimento, por meio da adição de metadados em algum nível do documento (palavra, frase, parágrafo). Este método, por ser essencialmente manual, requer a intervenção do usuário em todas as etapas, desde a seleção dos elementos do texto até a criação dos metadados.

Por se tratar de um processo manual, em alguns casos pode não ser a melhor abordagem, dependendo de variáveis como tamanho do *corpus* ou da base de documentos a serem anotados e do número de usuários realizando a tarefa. No entanto, é mais preciso se comparado com o método automático, para realização de anotações.

Anotação semântica semi-automática, requer, em algum ponto do processo, a intervenção humana. As aplicações nessa categoria podem diferir quanto à sua arquitetura, aos métodos utilizados para a extração de informação, à quantidade de emprego manual para que seja realizada a anotação e até mesmo quanto ao desempenho para realização de todo o trabalho.

Anotação semântica automática, na qual as aplicações que automatizam o processo de anotação em documentos Web se encontram, necessita da utilização de técnicas da IA (Inteligência Artificial), como aprendizado de máquina, para que aconteça. Além disso, algoritmos estatísticos podem ser utilizados para auxiliar o processo de anotação automática em imagens e vídeos. Embora a automatização do processo possa poupar esforço manual para a realização das anotações, pode não ser totalmente confiável para determinados tipos de aplicações, os documentos anotados não passem por algum tipo de revisão humana.

Além das categorias para as aplicações de anotação semântica, há uma classificação referente à estrutura das anotações (ANDREWS; ZAIHRAYEU; PANE, 2011) mapeando

os termos *tags*[8], atributos, relações e ontologias em um espectro, no qual o primeiro dos termos representa a forma mais fácil de anotação, e o último - ontologias - representa a mais difícil, do ponto de vista do usuário que realiza a anotação.

Uma *tag*, quando utilizada para anotar um recurso, descreve uma propriedade particular deste recurso, porém não carrega a semântica necessária para que, tanto máquinas quanto pessoas, entendam com clareza e sem ambiguidade o significado da propriedade descrita. Exemplos de usos de *tags* são comuns e utilizados em diversos serviços disponíveis na Web, como as *hashtags*, no Twitter. Sequência de caracteres precedida pelo caractere "`#`" utilizados para agrupar sintaticamente certa quantidade de mensagens, as *hashtags* não trazem informação semântica que permita agrupar as mensagens por significado. Mesmo tendo limitações e sendo relativamente simples para ser utilizada, esta estrutura não deixa de ter tanta importância que as outras.

Para a implementação de um modelo de anotação por *tags* não se faz necessária a utilização de um modelo baseado em triplas (Sujeito, Predicado e Objeto), como o RDF.

A anotação por meio de atributos utiliza a estrutura de chave e valor (*{key: value} pair*), na qual a chave representa a propriedade do recurso sendo descrito, seguida do seu valor. Ao contrário das *tags*, os atributos descrevem explicitamente as propriedades dos recursos e seus respectivos valores permitindo, desse modo, que outros tipos de aplicações sejam implementados como, por exemplo, uma busca por fotos em que a propriedade `resolução` com o valor `4290x2800 pixels` seja utilizada como filtro para os resultados.

Os elementos da anotação por meio de relações seguem a mesma estrutura do conjunto chave e valor da anotação por atributos, contudo os pares são formados pelos itens relação e recurso, *<Rel, Res>*. A relação *Rel* define como o recurso anotado é relacionado com o recurso *Res*. Anotações por relações podem ser utilizadas para definir relações entre elementos de um determinado recurso, e não somente entre recursos distintos.

Verificou-se, portanto, que anotações por ontologias representam o modelo mais rico, do ponto de vista estrutural, entre as outras formas apresentadas. Consiste na associação de um recurso ou suas partes, com a descrição de algumas de suas propriedades, respeitando um modelo conceitual formal ou uma ontologia. Esse modo de anotações permite ao usuário descrever e interligar recursos, qualificando-os como conceitos ou instâncias de conceitos e definindo as relações, propriedades e restrições existente entre eles.

[8] *Tag* pode ser traduzido como "etiqueta", no entanto, pelo contexto optou-se por manter o termo em inglês

2 Estado da arte

Este capítulo apresenta trabalhos e aplicações desenvolvidos e publicados recentemente e que sãos relacionados ao tema pesquisado. São apresentados exemplos de aplicações, utilizando tecnologias e conceitos de Web Semântica, além de trabalhos relacionados, com anotações semânticas.

2.1 *Google Structured Data*

A abordagem do Google para utilização de dados estruturados contribui para o aumento da base de conhecimento, chamada *Knowledge Graph*, com a inclusão de novos fatos, a partir dos dados estruturados, obtidos pelo robô de busca.

A Marcação de dados estruturados - ou *Structured data markup* - segue uma proposta padrão para anotação de conteúdo na Web a fim de tornar os dados *machine-readable*.

As ferramentas de busca Google, Yahoo, Microsoft Bing e Yandex utilizam os dados estruturados, obtidos das páginas indexadas, para apresentar resultados de busca precisos e, em alguns casos, experiências diferentes, como busca por voz.

O Google possui duas categorias de apresentação de conteúdos, obtidos a partir de dados estruturados: apresentação aprimorada nos resultados de busca (*Enhanced Presentation in Search Results*), e respostas fornecidas através do Knowledge Graph (*Answers from the Knowledge Graph*).

Google Rich Snippets

Ao fornecer dados estruturados juntamente com o conteúdo das páginas web, os algoritmos de busca do Google podem melhor indexar e interpretar o conteúdo coletado. Os resultados de busca do Google podem exibir os blocos de conteúdo chamados *Rich Snippets*. O conteúdo exibido nestes blocos são obtidos pelos algoritmos de indexação, a partir da leitura dos documentos web, nos quais são encontrados dados estruturados, que utilizam determinados vocabulários para descrevê-los.

Utilizando os vocabulários disponíveis no site Schema.org[1] para descrever e associar semântica aos dados, é possível prover informações para que os algoritmos de busca, tanto do Google quanto das outras empresas que fazem parte da iniciativa Schema.org, exibam resultados de pesquisas. No caso do Google, os *rich snippets* podem ser montados

[1] <http://schema.org>

conforme as seguintes categorias, a depender do vocabulário escolhido para descrever os dados:

Produtos

Os resultados com *rich snippets*, contendo produtos podem ser exibidos de dois modos: um bloco exibindo uma oferta específica de determinado produto ou um bloco exibindo ofertas agregadas do mesmo produtos em diferentes lojas, exibindo os preços mínimo e máximo.

O vocabulário utilizado para descrever dados de produtos é o <http://schema.org/Product>.

A figura 7 mostra um exemplo de utilização de vocabulários *Schema.org* nos resultados de busca do Google:

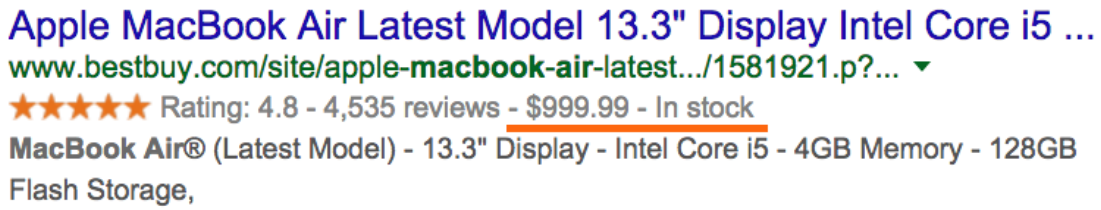


Figura 7: Resultado de busca do Google exibido com informações obtidas a partir dos dados estruturados

O código exibido abaixo descreve um produto de acordo com as propriedades do vocabulário *schema:Product*. São utilizados os atributos RDFa[2] no código HTML.

```
<div vocab="http://schema.org/" typeof="Product">
  <span property="brand">ACME</span>
  <span property="name">Executive Anvil</span>
  <img propertyu="image" src="anvil_executive.jpg"
  alt="Executive Anvil logo" />
  <span property="description">Sleeker than ACME's
  Classic Anvil, the
    Executive Anvil is perfect for the business traveler
    looking for something to drop from a height.
  </span>
  Product #: <span property="mpn">925872</span>
  <span property="aggregateRating" typeof="AggregateRating">
    <span property="ratingValue">4.4</span>
```

[2] <http://www.w3.org/TR/xhtml-rdfa-primer/>

```
      stars, based on <span property="reviewCount">89
        </span> reviews
    </span>

    <span property="offers" typeof="Offer">
      Regular price: $179.99
      <meta property="priceCurrency" content="USD" />
      $<span property="price">119.99</span>
      (Sale ends <time property="priceValidUntil" datetime="2020-11-05">
        5 November!</time>)
      Available from: <span property="seller" typeof="Organization">
                        <span property="name">Executive Objects</span>
                      </span>
      Condition: <link property="itemCondition"
      href="http://schema.org/UsedCondition"/>Previously owned,
        in excellent condition
      <link property="availability"
      href="http://schema.org/InStock"/>In stock! Order now!</span>
    </span>
  </div>
```

O trecho de código exibido acima e a documentação referente à utilização do vocabulário *schema:Product* pelo Google estão disponíveis no endereço <https://developers.google.com/structured-data/rich-snippets/products>.

Resenhas e Classificações

Avaliações sobre itens podem ser feitas por meio de resenhas (*Reviews*) e classificações (*Rating*): as resenhas são avaliações textuais, enquanto as classificações podem fornecer uma nota, em uma escala de 1 à 5.

O exemplo mostrado na 7 utiliza, além do vocabulário *schema:Product*, o vocabulário *schema:Review* para descrever os dados. As estrelas, o *rating* e os *reviews* exibidos no resultado de busca foram obtidos a partir dos dados estruturados, anotados com o vocabulário *schema:Review*.

Receitas O vocabulário *schema:Recipe* fornece algumas propriedades para descrever receitas culinárias presentes em determinadas páginas web.

A figura 8 mostra um resultado de busca para uma página contendo uma receita e fornece informações como o tempo total de cozimento do prato, além de informações sobre os *reviews* da receita.

Figura 8: Resultado de busca do Google para página contendo uma receita e descrita com uso do vocabulário *schema:Recipe*

Informações sobre o uso do vocabulário *schema:Recipes* para formar resultados do *Google Rich Snippets* estão disponíveis na página <https://developers.google.com/structured-data/rich-snippets/recipes>.

Eventos As marcações de eventos descrevem seus detalhes, como o tipo (festival de arte, festival de música, etc), a data de início e término, o local, além de informações sobre valores dos ingressos.

A lista de propriedades genéricas do vocabulário utilizado para descrever eventos está disponível em <http://schema.org/Event>. Há diferentes tipos de eventos que podem ser descritos com o uso dos vocabulários disponíveis no *Schema.org*, como eventos esportivos, festivais de arte, evento de negócios, entre outros.

A figura 9 exibe uma lista de eventos relacionados com o resultado de busca. Ela foi gerada a partir dos dados estruturados anotados utilizando os vocabulários do Schema.org.

Figura 9: Lista de eventos aparecendo em um resultado da busca do Google

Aplicativos Quando há dados sobre softwares e aplicativos marcados utilizando vocabulários do *Schema.org* as ferramentas do Google conseguem identificar precisamente estas informações e exibi-las nos resultados de busca, como mostra a figura 10.

A classe do Schema.org, que provê os atributos necessários para descrever o aplicativo, é chamada *SoftwareApplication* e se situa, de acordo com a organização hierárquica das classes, abaixo de *CreativeWork*.

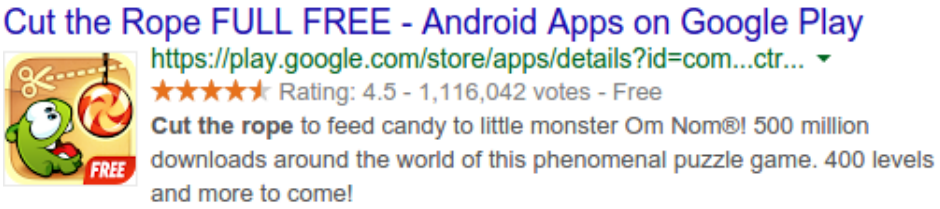


Figura 10: Resultado de busca sobre aplicativo para Android gerado a partir dos dados estruturados

A documentação referente ao uso do vocabulário *SoftwareApplication* para gerar itens do *Google Rich Snippets* está disponível em <https://developers.google.com/structured-data/rich-snippets/sw-app>.

Vídeos Os conteúdos em video, disponíveis nas páginas, podem ser descritos com as características do vocabulário *schema:VideoObject*, que fornece propriedades como *description*, *thumbnailUrl*, *uploadDate*, entre outras.

Figura 11: Lista de vídeos gerados a partir dos dados estruturados descritos com o vocabulário *schema:VideoObject*

Artigos As marcações presentes em artigos podem ser utilizadas nos resultados do Google, tanto na lista de notícias de determinado site quanto no bloco chamado “In the news”, conforme demonstrado nas figuras 12 e 13, respectivamente.

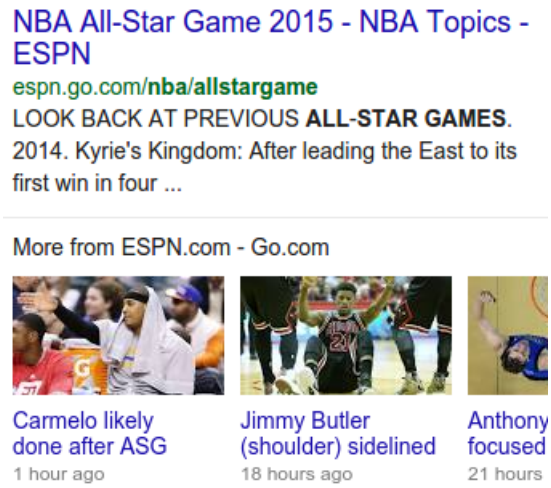


Figura 12: Lista de notícias do site ESPN.com montada a partir dos dados estruturados

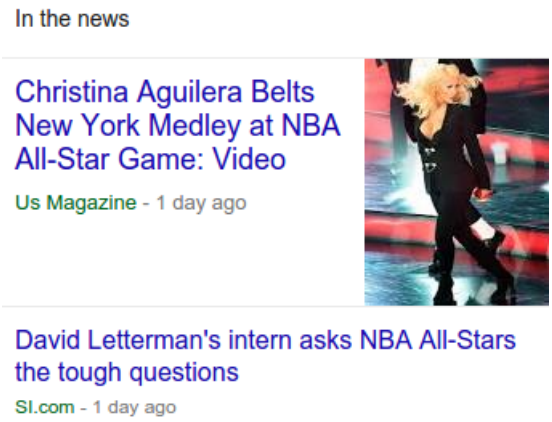


Figura 13: Bloco *In the news* com notícias geradas a partir dos dados estruturados

Para a anotação desses dados é utilizado o vocabulário *NewsArticle* do Schema.org. As propriedades utilizadas pelo Google para gerar os *Rich Snippets* de recursos anotados com *NewsArticle* estão disponíveis em <https://developers.google.com/structured-data/rich-snippets/articles>.

2.2 *Open Graph Protocol*

O *Open Graph Protocol* é uma iniciativa criada pelo Facebook para permitir que conteúdos externos ao "universo do Facebook" passem a fazer parte da rede, como um nó do grafo social.

O projeto consiste em um conjunto de termos padronizados, utilizados para descrever os recursos (páginas) disponíveis na Web.

A versão inicial do protocolo foi baseada na especificação RDFa (ADIDA et al., 2012) e, de acordo com o (RECORDON, 2010), as decisões técnicas em torno do projeto visavam tornar seu uso simples e prático.

O uso do OGP (*Open Graph Protocol*) ocorre pela adição dos termos do seu vocabulário às tags `<meta>` no cabeçalho de um documento HTML.

Para transformar uma página web externa em um objeto no grafo do Facebook é necessário utilizar, pelo menos, as quatro propriedades essenciais do vocabulário OGP, para fornecer os metadados básicos para o recurso:

og:title O título do objeto.

og:type O tipo do objeto, como por exemplo, "video.movie".

og:image A URL da imagem que representa o objeto.

og:url A URL do objeto. Esta URL será o ID permanente do objeto dentro do grafo.

Além dessas propriedades básicas, o vocabulário fornece propriedades para descrever imagem, vídeo, áudio e possui a descrição dos tipos de dados dos valores a serem utilizados para as propriedades (como *Boolean*, *Datetime*, *String*).

Ao fornecer este vocabulário para utilização em diferentes páginas web, o Facebook se torna capaz de obter os dados estruturados das páginas que seu *crawler* atingir; desse modo, as informações de determinadas páginas são exibidas como um objeto no grafo do Facebook, como no exemplo da figura 14:

Figura 14: Filme "*Iron Man*". Objeto gerado a partir de dados descritos com propriedades do *Open Graph Protocol*

Para que o filme "*Iron Man*" fosse representado como um objeto no grafo do Facebook foi necessário disponibilizar algumas informações sobre ele, descritas de modo estruturado com a utilização das propriedades do *Open Graph Protocol* nas tags meta no documento HTML:

```
<meta property="og:url" content="http://www.imdb.com/title/tt0371746/" />
<meta property="og:image" content="http://ia.media-imdb.com/images/M/
MV5BMTczNTI2ODUwOF5BMl5BanBnXkFtZTcwMTU0NTIzMw
&#064;&#064;._V1_UY1200_CR90,0,630,1200_AL_.jpg" />
<meta property="og:type" content="video.movie" />
<meta property="og:title" content="Iron Man (2008)" />
<meta property="og:site_name" content="IMDb" />
<meta property="og:description" content="Directed by Jon Favreau.
With Robert Downey Jr., Gwyneth Paltrow, Terrence Howard, Jeff Bridges.
After being held captive in an Afghan cave, an industrialist
creates a unique weaponized suit of armor to fight evil." />
```

Os exemplos representados pelo trecho de código HTML e pela figura 14 foram gerados a partir da página do filme "Iron Man", de 2008, disponível no IMDb com a URL <http://www.imdb.com/title/tt0371746/>.

A relação de termos definidos para o vocabulário do OGP estão disponíveis em <http://ogp.me/>.

O Facebook disponibiliza uma ferramenta na qual é possível realizar testes e verificar como os dados estruturados anotados com o vocabulário do *Open Graph Protocol* são obtidos das páginas. Esta ferramenta está disponível em <https://developers.facebook.com/tools/debug/og/object/>.

2.3 Suporte semântico à publicação de conteúdo jornalístico na Web

Dissertação de mestrado de autoria de Rafael Antônio Pinto Pena, defendida na PUC-Rio, sob orientação do Dr. Daniel Schwabe.

A pesquisa apresenta uma seção de estado da arte, que contempla casos de utilização de dados semânticos em empresas de mídia e traz a proposta de um modelo de publicação de conteúdo jornalístico na Web.

Na seção 2.2.1 do trabalho, é apresentada uma arquitetura projetada pelo IPTC (*International Press Telecommunication Council*) para facilitar o intercâmbio de notícias. A arquitetura faz uso de vocabulários para buscar atingir o objetivo do projeto em "criar

um ambiente que facilite aos usuários finais visualizarem conexões significativas entre os itens de notícias".

O trabalho propõe um modelo para publicação de conteúdo jornalístico na Web, que foi implementado na área de redação de notícias dos sites da companhia Globo.com. Juntamente com este modelo foi apresentado um conjunto de ontologias específicas utilizadas para dar suporte aos jornalistas, no momento da produção de conteúdo, para o site Globoesporte.com.

O fluxo para publicação de conteúdo proposto no trabalho possui dois objetivos principais, como sugere o autor (PENA; SCHWABE, 2012) "melhorar o processo de produção e viabilizar um novo tipo de oferta de conteúdos, de forma semiautomática, alavancado por modelos semânticos com suporte das tecnologias da Web Semântica".

A pesquisa realizada tem como proposta um modelo e uma ferramenta que faz uso de ontologias para facilitar a construção de narrativas jornalísticas para Web. A solução apresentada no trabalho é especializada, para utilização no domínio de futebol.

Como estudo de caso do trabalho foi apresentado um protótipo funcional da ferramenta, contendo a interface para autoria de matéria.

Nesta interface, desenvolvida com o auxílio do *plugin* de edição de texto *ckeditor*, o jornalista responsável por construir a matéria conta com o auxílio de informações da base de dados semantica, para a construção da narrativa.

A ferramenta permite ao jornalista inserir determinadas informações sobre os jogos, com base em dados coletados e armazenados pela equipe *Futpédia*. Ao escolher um confronto a partir da data de ocorrência, a ferramenta sugere alguns estereótipos com blocos de informações sobre o confronto.

Os estereótipos são gerados a partir da aplicação de ontologias, como na figura 15, mostrando a tela na qual o sistema sugere que o confronto possui o *template* de "jogo movimentado", pelo fato de os dados mostrarem isso.

A intervenção humana, neste caso, é apenas a escolha do jornalista em utilizar o bloco de informações - chamado *template*, na matéria ou não. Mas o *template* é gerado e sugerido automaticamente.

A sequência de imagens mostra o processo, desde a escolha do *template* à publicação da notícia, utilizando as informações do *template*:

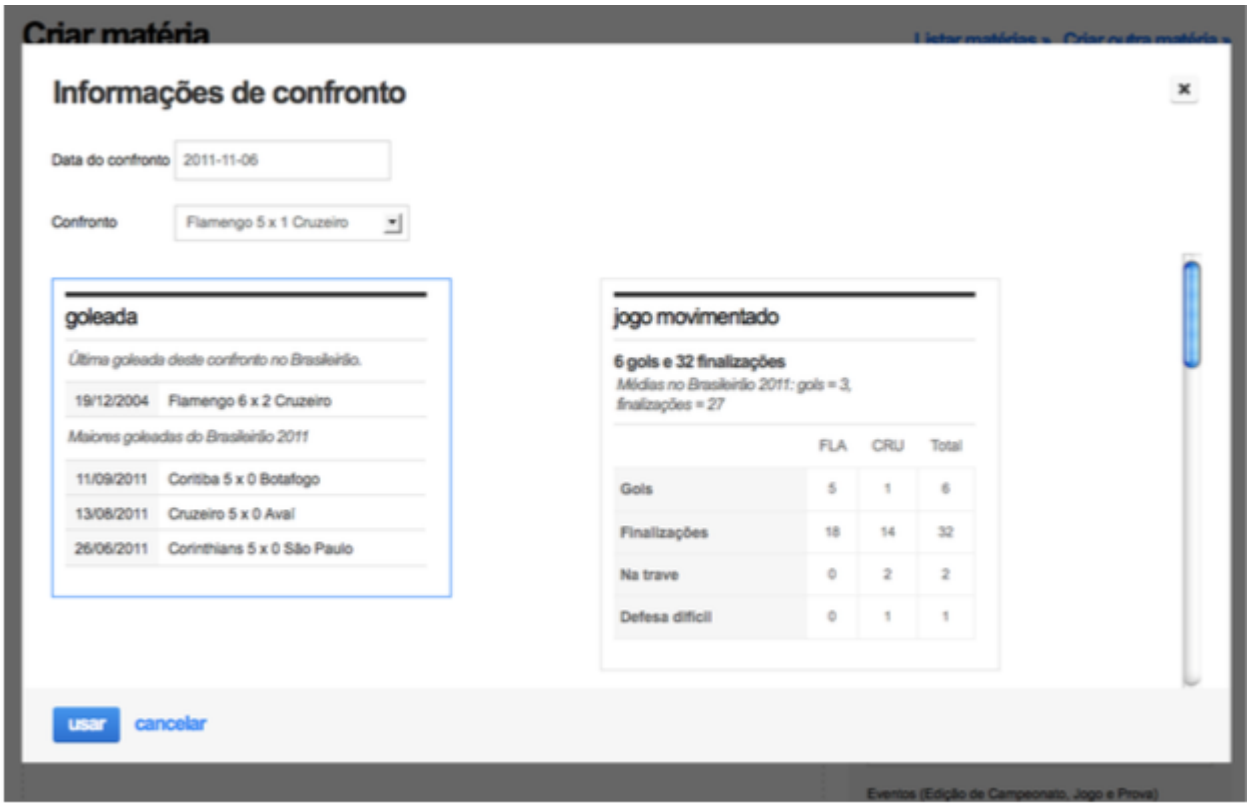


Figura 15: Imagem de um *template* para publicação de narrativas jornalísticas. Extraído de (PENA; SCHWABE, 2012)

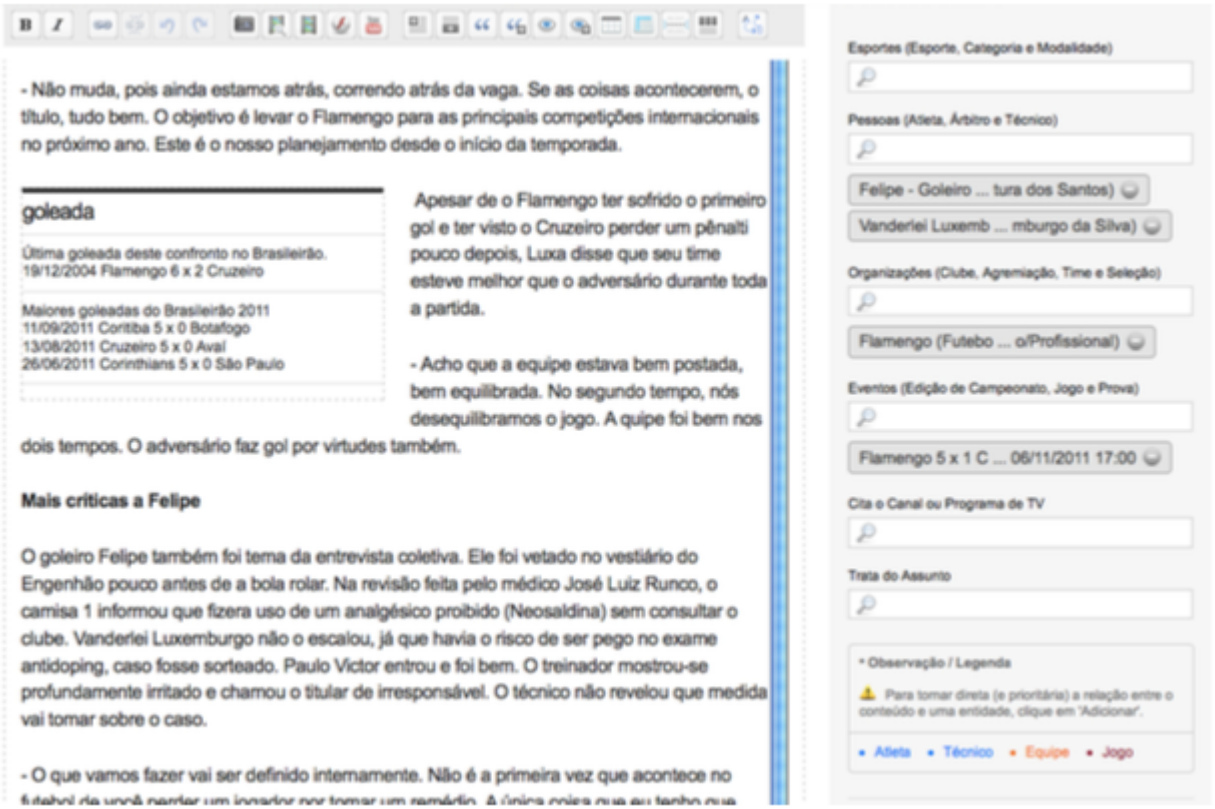


Figura 16: Interface de criação de matérias com a utilização de um *template*. Extraído de (PENA; SCHWABE, 2012)

06/11/2011 20h21 - Atualizado em 06/11/2011 22h32

Luxa não se empolga com goleada e mantém meta: vaga na Libertadores

Técnico volta a criticar goleiro Felipe e diz que vai avaliar caso internamente

Por **Richard Souza**
Rio de Janeiro

Tweetar 80 Recomendar 105

Vanderlei Luxemburgo não se deixou levar pela vitória empolgante do Flamengo. Neste domingo, o time goleou o Cruzeiro por 5 a 1 no Engenhão e voltou com força à briga pelo título brasileiro. Com 55 pontos, está a três do líder Corinthians, que foi derrotado pelo lanterna América-MG (2 a 1). O treinador fez questão de frisar que o discurso não vai mudar. A prioridade é, pelo menos por enquanto, a vaga na Libertadores da América.

- Não muda, pois ainda estamos atrás, correndo atrás da vaga. Se as coisas acontecerem, o título, tudo bem. O objetivo é levar o Flamengo para as principais competições internacionais no próximo ano. Este é o nosso planejamento desde o início da temporada.

goleada

Última goleada deste confronto no Brasileirão.

19/12/2004	Flamengo 6 x 2 Cruzeiro

Maiores goleadas do Brasileirão 2011

11/09/2011	Coritiba 5 x 0 Botafogo
13/08/2011	Cruzeiro 5 x 0 Avaí
26/06/2011	Corinthians 5 x 0 São Paulo

Apesar de o Flamengo ter sofrido o primeiro gol e ter visto o Cruzeiro perder um pênalti pouco depois, Luxa disse que seu time esteve melhor que o adversário durante toda a partida.

- Acho que a equipe estava bem postada, bem equilibrada. No segundo tempo, nós desequilibramos o jogo. A quipe foi bem nos dois tempos. O adversário faz gol por virtudes também.

Figura 17: Notícia publicada utilizando um *template* gerado automaticamente. Extraído de (PENA; SCHWABE, 2012)

O objetivo do trabalho foi verificar se um modelo semântico para produção de conteúdo, como o apresentado, poderia agregar valor ao modelo utilizado na organização onde a pesquisa foi realizada.

Segundo os autores (PENA; SCHWABE, 2012), o grupo de usuários que fez parte da amostra para validação dos resultados da pesquisa apontou que o modelo de publicação foi considerado inovador em relação ao existente na época e se pôde afirmar que a ferramenta influenciou não somente na forma, mas também no conteúdo da publicação.

2.4 *Thomson Reuters Open Calais*

Open Calais[3] é um produto da empresa *Thmoson Reuters* que faz a análise de textos, utilizando técnicas de inteligência artificial. como aprendizagem de máquina e processamento de linguagem natural, a fim de identificar entidades, eventos, relações, tópicos e *tags* presentes nos textos analisados.

O serviço disponibiliza uma API[4], que pode ser utiliza para fins comerciais, na qual é possível submeter conteúdo não estruturado e obter como resposta conteúdo estruturado contendo as entidades, empresas e indivíduos identificados no item analisado.

O *Open Calais* mapeia as *tags* de metadados em identificadores únicos, para auxiliar na desambiguação de conteúdo.

Informações das diferentes bases de dados da *Thomson Reuters* são utilizadas pelo *Open Calais* no suporte à identificação de entidades e no processamento dos textos. O serviço analisa automaticamente a entrada de dados e executa dois diferentes processos para obter informações relevantes sobre o conteúdo analisado: reconhecimento de entidades e relacionamentos (*Named Entity and Relationship Recognition*), e identificação dos assuntos do conteúdo (*Aboutness Tagging*), conforme apresentado no documento *Thomson Reuters Open Calais - API User Guide* (2015).

No processo de reconhecimento de entidades e relacionamentos, o *Open Calais* analisa o texto de entrada, buscando referências a companhias, indivíduos, cidades, produtos, ações de bolsas de valores, entre outros. Quando há menções diretas a esses itens, eles são classificados como *Entidades*. Referências que não são diretas, como recomendações de analistas e abertura de capital em bolsa de valores são classificadas como *Relações*.

Como resultado desse primeiro processo, o sistema retorna um conjunto de *tags* e marcações:

Instance Tags: cada referência encontrada pelo *Open Calais* é expressa como uma *Instance Tag*.

[3] <http://www.opencalais.com/>
[4] Application Programming Interface

Entity Markup Tag: Cada grupo de uma ou mais tags que referenciam para uma única entidade são expressas como *Entity Markup Tag*. Diferentes menções à mesma pessoa resultam em uma única *Entity Markup Tag* do tipo Pessoa.

Relevance Tag: Indica qual a relevância da entidade ou da relação identificada.

Confidence Tag: Indica quão segura é a afirmação sobre determinada entidade, em que grau de certeza uma entidade do tipo Pessoa se refere, de fato, a uma pessoa.

Disambiguation Tag: O *Open Calais* ttenta realizar um mapeamente da entidade ou relação extraído do documento para uma entidade com identificador único, presente no conjunto de dados da *Thomson Reuters*. Se esse mapeamento ocorre com sucesso, uma tag de desambiguação é gerada. Esse mapeamente permite que todas as instâncias que se referem à mesma "coisa" sejam identificadas, sem ambiguidades entre todos os documentos processados pelo *Open Calais*.

Como resultado do segundo processo - *Aboutness Tagging* - o Open Calais retorna metadados sobre o conteúdo como um todo e não sobre as entidades separadas, como no primeiro processo. As classificações para o documento são as seguintes:

Social Tagging: Classifica o documento de acordo com a *folksonomia*[5] da Wikipedia.

Category Tagging: Identifica os tópicos apresentados no documento com base nas taxonomias *Thomson Reuters Coding Schema* (TRCS) e International Pres Telecommunications Council (IPTC).

Industry Tagging: Identifica os setores e indústrias relacionados com o conteúdo analisado, de acordo com a taxonomia *Thomson Reuters Business Classification* (TRBC).

No site do *Open Calais*[6] existe uma página de demonstração na qual é possível adicionar algum conteúdo, submeter e obter os dados estruturados resultantes da análise do *Open Calais*. O formulário para demonstração está disponível em <http://www.opencalais.com/opencalais-demo/> (Acessado em 27/11/2015 às 10:21).

O uso do sistema *Open Calais* acontece por meio de chamadas na API disponibilizada pelo produto.

Para autenticação e utilização da API do *Open Calais* é necessário utilizar uma chave de acesso fornecida pela Thomson Reuters.

Os formatos de entrada permitidos para serem enviados ao sistema são `text/html`, `text/xml`, `text/raw`.

O sistema tem suporte para conteúdos escritos em Inglês, Francês e Espanhol.

[5] *Folksonomy*: maneira de classificar e indexar informações realizado por pessoas, de modo colaborativo.
[6] <http://www.opencalais.com>

2.5 *Epiphany*

Epiphany é um serviço para anotação de Dados Conectados (do inglês *Linked Data*) em páginas web, de modo automatizado. Qualquer conjunto de dados em RDF, no escopo do *Epiphany*, é considerado *Linked Data* e pode ser utilizado nos processos de anotação semântica do serviço (ADRIAN et al., 2010).

Com a extração de informações, com base na ontologia e no conjunto de dados conectados, o serviço gera um grafo RDF, com dados sobre o conteúdo da página a ser anotada. A partir desse grafo RDF são geradas as anotações que são integradas, utilizando RDFa, com o HTML do conteúdo analisado a fim de gerar uma versão da página com HTML+RDFa.

A utilização de RDFa permite aos *web authors* (como classificados por ADRIAN et al.) anotarem seus conteúdos com marcações semânticas e então conectar seus textos não estruturados com o mundo de dados *machine-readable*.

A aplicação analisa e reconhece instâncias e propriedades em uma página web, a partir de uma ontologia, e fornece anotações em RDFa, com a descrição de tais entidades reconhecidas no texto.

O serviço utiliza uma técnica chamada *Ontology-Based Information Extraction* (OBIE) para análise e reconhecimento do conteúdo das páginas.

Diferentemente do serviço *Open Calais* (descrito na seção 2.4), que trabalha com as ontologias da *Thomson Reuters*, especializadas ao domínio de notícias, o *Epiphany* pode utilizar ontologias de diferentes domínios para gerar as anotações.

Ao final do processo de análise do conteúdo, reconhecimento e extração de entidades, e geração da marcação semântica, uma página web obtida como *output* do *Epiphany* se parece com a página mostrada na figura 18:

Figura 18: Página web com HTML+RDFa gerada pelo Epiphany. Fonte: (ADRIAN et al., 2010)

Todos os termos contornados pela borda laranja presentes na página tiveram "sua semântica atribuída" pelo Epiphany, ou seja, há dados estruturados com a semântica desses termos disponíveis.

Ainda como visualização do conteúdo semântico, é possível clicar em algum dos conceitos e obter mais informações a seu respeito, conforme mostrado na figura 19:

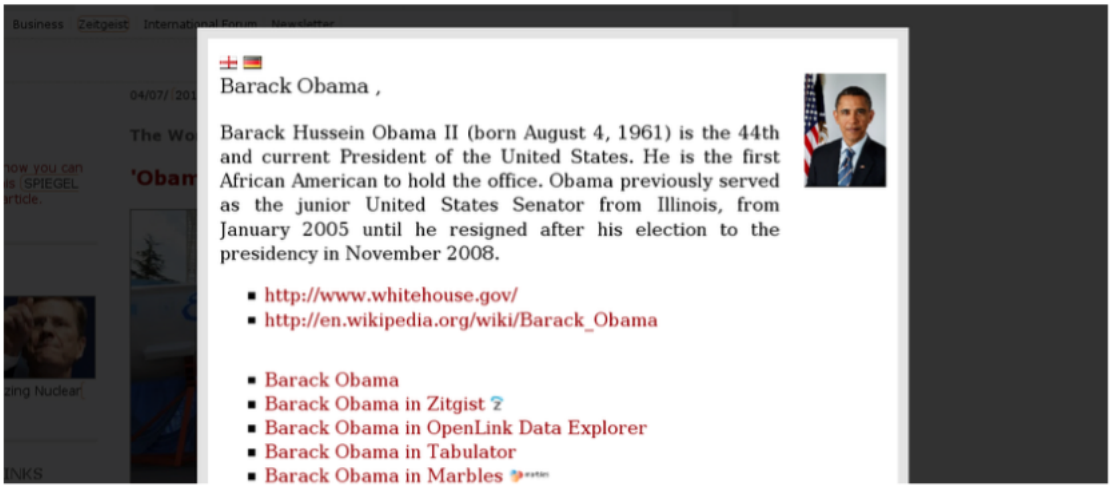


Figura 19: Bloco com mais informações sobre o termo anotado semanticamente pelo *Epiphany*. Fonte: (ADRIAN et al., 2010)

A abordagem utilizada no *Epiphany* emprega a técnica de anotação automática, com o auxílio de OBIE, para geração dos metadados sobre recursos encontrados em meio ao conteúdo das páginas analisadas.

3 Proposta de ferramenta para anotação semântica

Embora a área de Web Semântica tenha recebido bastante atenção no ambiente acadêmico e com isso tenha evoluído em direção à definição de padrões, são poucas as aplicações e inciativas desenvolvidas para atender as necessidades dos usuários finais, aponta KARGER (2014). Considerando o distanciamento entre os usuários finais e essas tecnologias, esta pesquisa investiga a viabilidade técnica de desenvolvimento de uma aplicação, com o propósito de contribuir com a participação de mais usuários finais no ecossistema de Web Semântica.

A ferramenta proposta permite que usuários publicadores de conteúdo na Web contribuam com o ecossistema de Web Semântica, a partir da geração de dados estruturados e metadados. Ao selecionarem palavras, frases ou parágrafos em seus textos e realizarem o processo de anotação semântica a fim de associar tais objetos selecionados às definições encontradas em ontologias e vocabulários compartilhados, esses usuários fornecerão semântica aos dados que estão sendo publicados.

3.1 Visão geral

A ferramenta proposta neste trabalho é uma aplicação de anotação semântica com foco em usuários publicadores de conteúdo na Web. O objetivo da aplicação é permitir que mais usuários passem a gerar conteúdos com dados estruturados de modo que possam ser processados por máquinas (conteúdo *machine-readable*).

Assume-se como premissa para a utilização da aplicação que o usuário possua familiaridade com o conceito de Web Semântica e tenha capacidade de escolher, dentre uma lista pré-definida, a ontologia adequada ao domínio dos dados a serem anotados. Isso, contudo, não requer que o usuário possua conhecimentos mais aprofundados em relação à Web Semântica, como construção de ontologias, domínio das linguagens *Turtle*, *OWL*, entre outras. Desse modo, portanto, acredita-se diminuir as barreiras de utilização de uma aplicação de Web Semântica.

A aplicação contempla dois estágios do que se classifica como “Dados na Web” e abrange tanto a etapa de publicação quanto a de consumo dos dados.

A etapa de publicação consiste da utilização da ferramenta por parte do usuário, para a realização de anotações no conteúdo que virá a ser publicado. As anotações são realizadas por meio da atribuição de valores às propriedades de determinados vocabulá-

rios compartilhados, a fim de descreverem o sujeito que está sendo anotado. Ao final do processo de anotação podem ser obtidas triplas RDF, com as respectivas descrições dos dados.

Os grafos RDF gerados a partir das anotações são armazenados em uma base de dados adequada, como o *Open Link Virtuoso* e são posteriormente utilizados na etapa de consumo, gerando informações sobre os recursos anotados.

Com o propósito de ilustrar esta breve apresentação da ferramenta, consideremos o seguinte caso como exemplo:

Em determinada organização são realizadas reuniões periódicas a fim de discutir temas pertinentes ao seu funcionamento.
Todas as reuniões são devidamente acompanhadas e transcritas para posterior consulta; no entanto, as pautas geradas são escritas em linguagem natural e sendo adequadas para a interpretação por humanos, ou seja, o conteúdo gerado é considerado *human-readable*. Surge a necessidade de levantar quais foram os temas discutidos e quais eram os membros participantes das reuniões; porém, as buscas no conteúdo *human-readable*, pode não trazer resultados tão precisos, além de que o esforço para realizar esse levantamento de modo manual pode ser massante, considerando que possam existir um grande número de reuniões.

Embora existam diferentes maneiras para se fazer tal levantamento de informações, uma abordagem sugerida poderia levar em conta aspectos de Web Semântica para obter tais respostas. Devemos considerar que os dados das pautas das reuniões são estruturados (*machine-readable*) com a utilização de ontologias, tanto para descrever os assuntos discutidos quanto para identificar quais eram as pessoas que estavam presentes.

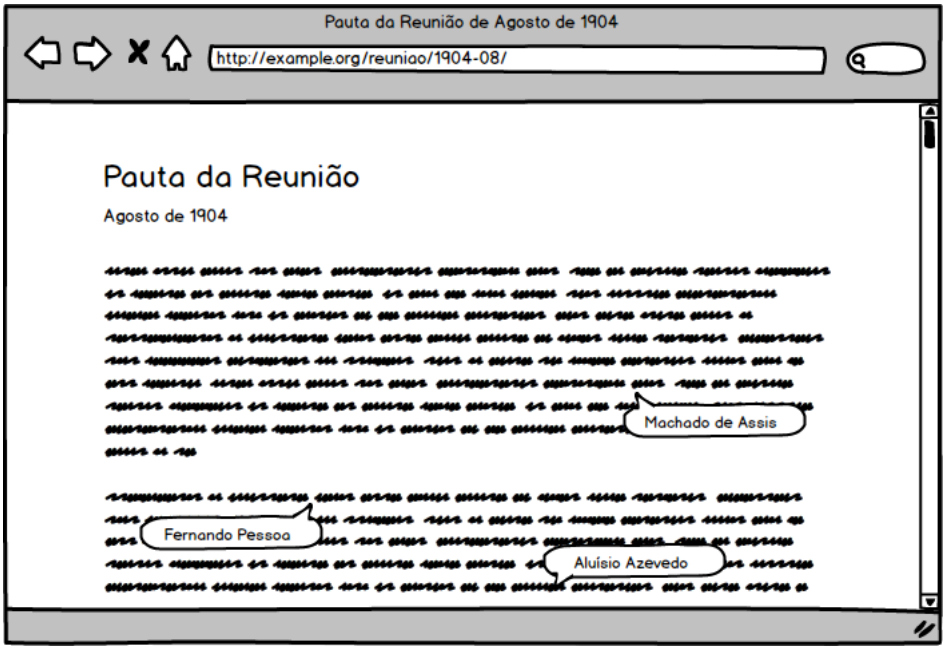


Figura 20: Ilustração mostrando uma página com a ata de uma reunião.

Os membros presentes nas reuniões poderiam ser identificados nos textos das pautas com o uso de uma ontologia específica para descrever pessoas como, por exemplo, o vocabulário *FOAF*. Desse modo, as cadeias de caracteres dos respectivos nomes dos membros, quando anotadas com o vocabulário *FOAF*, poderiam ser interpretadas pela máquina como recursos do tipo Pessoa.

Este é um dos inúmeros casos nos quais podem ser utilizadas tecnologias relacionadas com Web Semântica para obter resultados satisfatórios. E a ferramenta proposta neste trabalho visa possibilitar tais aplicações, como descrita no caso hipotético.

A etapa de consumo ocorre a partir do momento em que os dados estruturados tenham sido armazenados, seguindo o modelo RDF, com sujeito, predicado e objeto, e que cada recurso anotado semanticamente possua um identificador único, um URI.

Passa a ser possível construir páginas web com informações obtidas a partir dos grafos RDF, gerados na etapa de publicação. Novamente tomaremos como exemplo uma aplicação sobre o caso hipotético já descrito:

Todas as pautas das reuniões foram anotadas semanticamente e possuem, portanto, dados estruturados sobre seus conteúdos, além de identificadores únicos para que possam ser acessadas na Web. A página de uma pauta poderia exibir, além do conteúdo textual em formato *human-readable*, um bloco com os nomes dos membros que participaram da discussão, obtendo esses dados diretamente do grafo RDF.

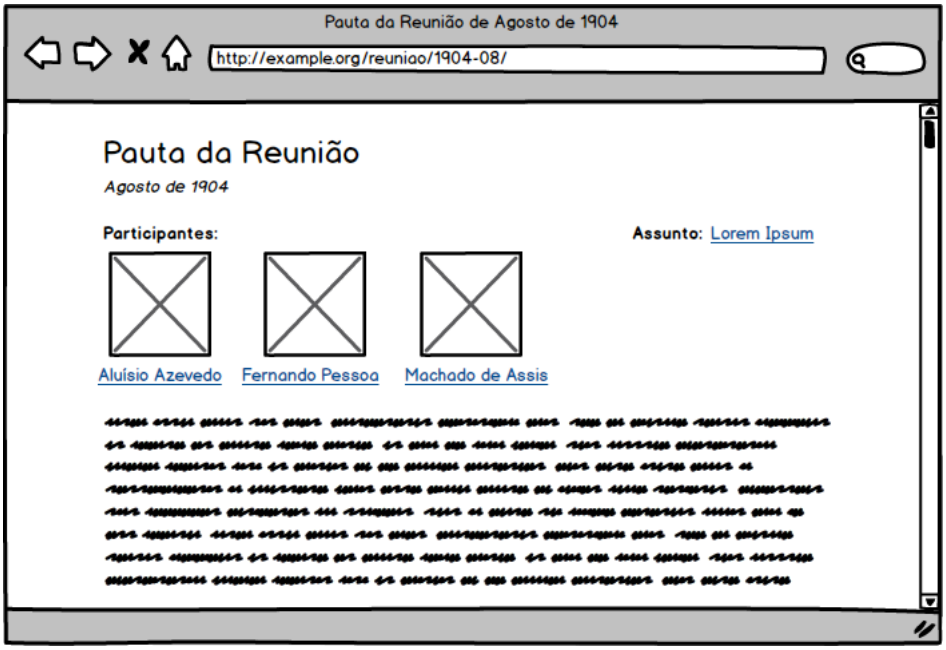


Figura 21: Ilustração mostrando uma página com a ata e os participantes de uma reunião.

Além das pautas, outros recursos anotados semanticamente também recebem um URI; desse modo, podem existir páginas para os membros participantes das reuniões. Quando a página de um membro específico é acessada, podem ser exibidas suas informações pessoais e de contato, mas também uma lista com todas as reuniões que tenha participado e os temas de cada uma delas. Essas informações sobre as reuniões podem ser obtidas do grafo RDF, gerado pelas anotações.

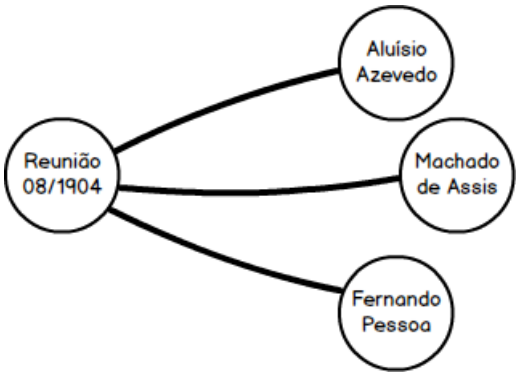


Figura 22: Ilustração com um grafo representando alguns dados sobre reunião.

O modelo proposto para a ferramenta é de uma aplicação que funcione em um editor de texto na Web, no qual este (usuário publicador de conteúdo) possa escolher, dentre uma lista pré-definida de vocabulários, quais utilizará para realizar descrição dos dados, em meio ao conteúdo a ser publicado.

Os vocabulários e ontologias disponíveis na aplicação são adicionados previamente, através de uma tela específica da aplicação. Cada vocabulário a ser utilizado na aplicação

possui uma cor associada, esta cor identificará os conceitos do texto, anotados com o respectivo vocabulário.

Cabe ao usuário editor de conteúdo escolher tanto o vocabulário quanto as propriedades adequadas para a descrição dos dados presentes no texto.

No momento em que o usuário seleciona parte do conteúdo, podendo ser uma única palavra ou uma frase do texto, a ferramenta exibe uma caixa com opções de ação e uma das opções exibidas é a caixa de texto, na qual o usuário define qual o tipo do conceito que será anotado (tipo *Person*, *Organization*, *Book*, etc), de acordo com o vocabulário adequado.

Após a escolha do tipo do conceito anotado, é necessária a definição de um identificador único para aquele recurso, um URI. Então, uma caixa de texto é exibida para que o usuário escolha um URI já definido anteriormente ou permita que a aplicação forneça o URI.

Concluído o primeiro passo, que consiste na definição de um identificador e do tipo do recurso, as primeiras triplas RDF são geradas pela aplicação:

```
<http://example.org/Person/Recurso/> rdfs:type foaf:Person ;
<http://example.org/Person/Recurso/> rdfs:label ‘‘Texto anotado’’ ;
```

Este mesmo recurso pode vir a ser utilizado em outros textos, por isso a necessidade de existir um identificador único para que seja referenciado em diferentes páginas.

Na mesma caixa de opções que surgiu para o conteúdo selecionado, é possível que o usuário utilize outros predicados para definir novos valores para o recurso. Nela pode ser definido, por exemplo, o endereço de e-mail do recurso anotado por meio da propriedade `foaf:mbox`. Assim a aplicação gera mais uma tripla para o sujeito `<http://example.org/Person/Rec`

```
<http://example.org/Person/Recurso/> foaf:mbox ‘‘email@example.org’’
```

Podem ser adicionadas quantas propriedades forem necessárias para a descrição do recurso.

Ao final desse processo de anotação, o conteúdo anotado aparece no texto com uma borda e cor de fundo correspondentes à cor da ontologia que foi utilizada para definir o seu tipo, ou seja, da cor do vocabulário do conceito definido em `rdfs:type` para o referido sujeito.

O editor é o usuário responsável por escolher quantos e quais conceitos serão anotados semanticamente no conteúdo a ser publicado.

Todo o conteúdo estruturado gerado pela aplicação é armazenado em bases de dados adequadas, como bancos de dados para armazenamento de grafos RDF e esse conteúdo pode ser utilizado em diferentes aplicações, como mostrado no exemplo das pautas de reuniões.

A aplicação possibilita que usuários publicadores de conteúdo disseminem dados estruturados na Web sem, de fato, terem que escrever uma tripla RDF em linguagem *Turtle*, por exemplo. Isso permite que mais usuários finais possam contribuir com o ecossistema de Web Semântica, sem necessariamente possuir conhecimentos específicos sobre o assunto.

3.2 Descrição técnica

Esta seção expõe os aspectos técnicos e arquiteturais da aplicação base para a ferramenta proposta no trabalho.

Os componentes da ferramenta são organizados de acordo com o padrão arquitetural *MVC* (*Model-View-Controller*), amplamente adotado pela comunidade de Engenharia de Software (SILVA, 2012).

Na camada *Model* desta arquitetura, encontram-se os bancos de dados específicos para cada demanda da aplicação, como o banco de dados relacional para o conteúdo e o banco de dados para armazenamento de estruturas modeladas em grafos, utilizado para armazenar os dados em RDF.

Outros componentes da arquitetura, como os gerenciadores de URI e de Vocabulários, estão estruturados na camada *Controller* e os componentes que realizam a interface com o usuário, como o formulário de anotação, estão na camada *View*.

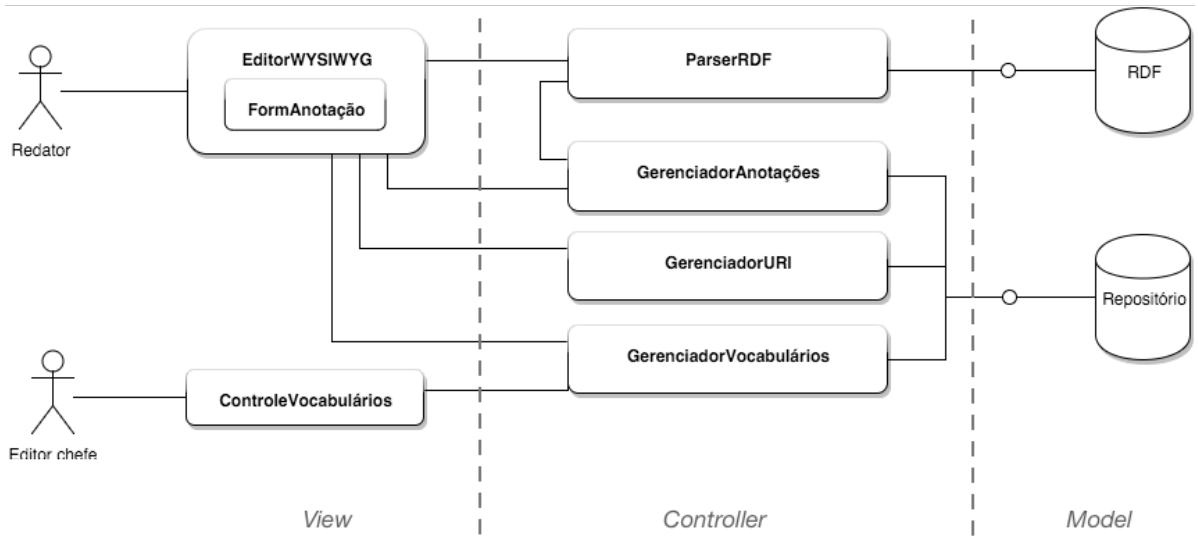


Figura 23: Esquema com os elementos que formam arquitetura da aplicação

A descrição dos componentes que formam a arquitetura da aplicação será apresentada nesta seção. Além das restrições para cada componente, são apresentadas sugestões de tecnologias para implementação dos mesmos.

3.2.1 Componente de gerenciamento de URI

O componente de gerenciamento de URI passa a funcionar a partir do momento em que o usuário seleciona um trecho do conteúdo para realizar a anotação e atribui o tipo do conceito anotado. Este componente é responsável por garantir que cada recurso anotado tenha um identificador único correspondente.

Para a formação do identificador do recurso que está sendo anotado, o componente de gerenciamento de URI recebe como dados de entrada dois parâmetros: o tipo do conceito e o texto selecionado.

Como "tipo do conceito" espera-se o valor correspondente à propriedade *rdfs:type* do recurso, e o "texto selecionado" refere-se ao conteúdo que será anotado.

O tipo do conceito é utilizado na formação do URI para evitar ambiguidades entre dois recursos de tipos diferentes, porém com nomes ou rótulos iguais, por exemplo:

```
<http://example.org/Organization/Apple> rdfs:type schema:Organization
<http://example.org/Fruit/Apple> rdfs:type frt:Fruit
```

Neste exemplo, o primeiro recurso refere-se à empresa *Apple*, enquanto o segundo à fruta maçã (no caso, *Apple*, em inglês).

O parâmetro "texto selecionado" também é utilizado na formação do identificador, mas antes de fazer parte do URI o parâmetro passa por um processo de adequação para remoção de determinados caracteres, espaços em branco, etc. O método para adequação do texto consiste nos seguintes passos:

- Remoção dos espaços em branco no início e fim do texto. Realizado com o método *trim()*, presente em muitas linguagens de programação.
- Divisão do conjunto das palavras presentes no texto por meio de um processo chamado *Tokenization*. Ao fim deste processo é obtida uma lista de *tokens*, cada *token* representa uma palavra.
- A cadeia de caracteres formada para o URI segue a mesma convenção adotada nos URI da Wikipedia, com o caractere "_" (*underscore*) como separador entre as palavras. Desse modo, os *tokens* são mesclados, utilizando o caractere "_" no lugar dos espaços em branco.

- Os caracteres "#" (cerquilha) e "." (ponto) são omitidos na formação da URI.
- O URI formado, de acordo com o esquema "/Tipo/Recurso/'', é salvo juntamente com o *trim("texto selecionado")* em um repositório com todos os identificadores gerados para os recursos.
- Como resultado, este método deve retornar o URI formado para que seja utilizado em outros componentes do sistema.

Além da responsabilidade pela formação de novos identificadores, o componente de gerenciamento de URI possui um mecanismo para consultar, quando necessário, os URI existentes.

Determinado trecho do texto pode se referir a um recurso já anotado e, consequentemente, com um identificador. Em casos como esse, é interessante que exista a possibilidade de pesquisar entre os URI formados qual se refere ao recurso apontado.

Continuando com um exemplo já descrito acima, no qual a organização *Apple* foi anotada semanticamente e possui o identificador `http://example.org/Organization/Apple`. A empresa é citada em um outro texto com conteúdo sobre mercado de ações, mas ao invés de aparecer o nome "*Apple*", no texto aparece o termo *AAPL*, que se refere ao código do papel da *Apple* negociado na bolsa de valores Nasdaq. Para anotar o conteúdo, não será necessária a geração de um novo URI, considerando que o termo se refere ao recurso "*Apple*" apontado previamente. Então, neste caso, deve ser utilizada a opção de busca de indicadores que aparece na caixa de opções ao selecionar algum trecho do texto. Os parâmetros da busca podem ser tanto partes de uma URI existente, quanto partes de um texto previamente anotado. A tripla RDF formada após esta anotação possui a seguinte estrutura:

```
<http://example.org/Organization/Apple> rdfs:label ''AAPL''
```

Quando são anotados novos conteúdos a respeito do mesmo recurso, o repositório de URI deve ser atualizado, indicando qual novo rótulo determinado indicador recebeu.

O comportamento esperado para este componente é representado nos diagramas das figuras 24 e 25.

Um protótipo da ferramenta vem sendo desenvolvido paralelamente à pesquisa e foram utilizadas, nesse protótipo, algumas linguagens e bibliotecas para implementar o componente de gerenciamento de URI.

Para o repositório de URI está sendo testado o banco de dados *MongoDB*, com o conector específico para a linguagem PHP. O elemento responsável por validar as novas URI e adicioná-las ao banco é desenvolvido em PHP. A interface gráfica pela qual o usuário

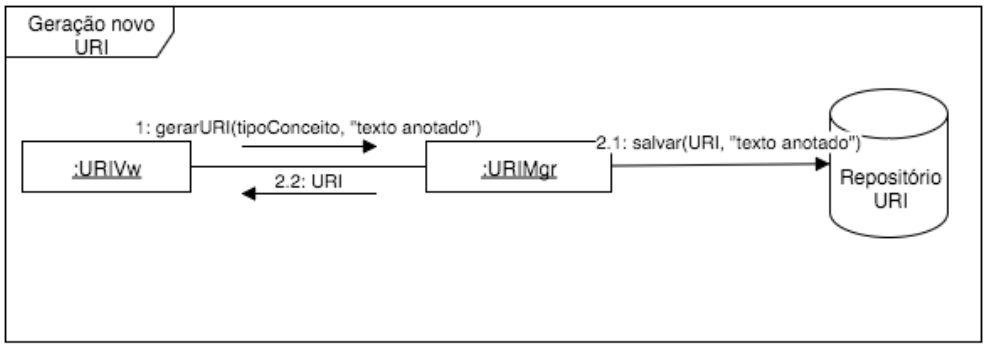


Figura 24: Geração de um URI novo

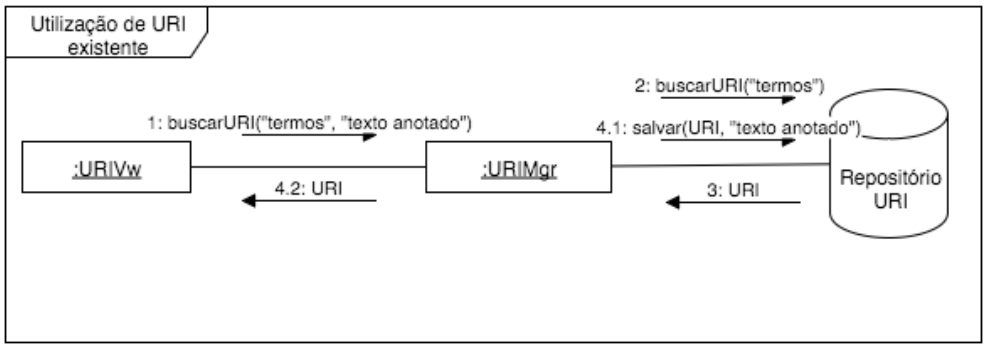


Figura 25: Busca e atualização de um identificador na base de dados

interage com a aplicação é desenvolvida em *Javascript*, com a utilização da biblioteca de código aberto, chamada *Medium-editor*[1], disponível no site *Github*.

3.2.2 Componente de gerenciamento de vocabulários

A inserção de vocabulários no banco de dados da aplicação é realizada através do componente de gerenciamento de vocabulários, responsável pelas ações de adição, exclusão, listagem e recuperação[2] das ontologias e vocabulários utilizados na ferramenta.

A adição ao repositório acontece a partir do momento em que o usuário fornece o URI de um vocabulário disponível na Web para que um componente específico realize o processo de *parsing* do conteúdo, a fim de obter todos os termos disponíveis no vocabulário, incluindo as classes e as propriedades. O usuário deve especificar um prefixo para identificação e utilização do vocabulário na aplicação.

O URI do *namespace* e o prefixo são os parâmetros de entrada para o método que realiza a análise do conteúdo, para a extração dos termos do vocabulário. As classes e propriedades obtidas do conteúdo do arquivo são organizadas em uma estrutura de dados que é posteriormente armazenada no repositório.

[1] <https://github.com/yabwe/medium-editor>
[2] A palavra “recuperação” foi utilizada com a mesma conotação do termo inglês *retrieval*.

Alguns vocabulários possuem URI específico para uma página contendo informações *human-readable*, como descrição dos termos, dados do responsável pela disponibilização, versão, etc; e um URI específico para o arquivo *machine-redable* do vocabulário, chamado de *namespace*. O *namespace* contém as classes e termos do vocabulário em formato *machine-readable*, como *RDF/XML* ou *Turtle*.

Outros vocabulários possuem o mesmo URI tanto para o conteúdo *human-readable* e *machine-readable* e os arquivos são servidos em diferentes versões, utilizando a técnica de negociação de conteúdo (do inglês *Content-Negotiation*), definida na especificação HTTP e implementada em servidores web, como o *Apache*.

Para o repositório de vocabulários é necessário manter, pelo menos, a URI do *namespace* do vocabulário.

São necessárias, além dos dois parâmetros de entrada, outras informações sobre o vocabulário que podem ser preenchidas por meio de um formulário. O nome e a descrição do vocabulário devem ser fornecidos, além de haver a escolha de uma cor que será utilizada para identificação do vocabulário e dos conceitos anotados no texto.

Os dados relacionados a cada vocabulário são organizados conforme a estrutura mostrada na figura 26, utilizando o formato *JSON*.

A estrutura escolhida para organizar os dados de um vocabulário visa facilitar a utilização do mesmo e de seus termos no processo de anotação de conteúdo, realizado na camada *View*.

Ao optar por determinado vocabulário para a descrição dos termos, a aplicação completa automaticamente tanto o prefixo quanto os termos do vocabulário a serem utilizados. Quando um prefixo é escolhido, toda a estrutura de dados referente àquele vocabulário é carregada do repositório e os nomes dos termos possíveis de serem inseridos no campo de predicados são limitados ao conteúdo da lista armazenada na propriedade *terms*. Isso visa facilitar a implementação do comportamento chamado *autocomplete*, para auxiliar o usuário no momento de realização da anotação.

O gerenciamento do repositório de vocabulário deve ser realizado em uma página separada da página com o editor de texto para as anotações. Na página são disponibilizados os métodos de adição, edição e exclusão dos vocabulários.

O banco de dados *MongoDB*, o mesmo utilizado para o repositório de URI, está sendo utilizado para o repositório de vocabulários.

3.2.3 Componente de gerenciamento de anotações

Cada anotação realizada por meio da ferramenta deve ser unicamente identificada, com o propósito de reconhecer a anotação como um nó exclusivo no grafo.

```
{
   "uri":"http://www.w3.org/TR/vocab-org/",
   "namespace":"http://www.w3.org/ns/org#",
   "prefix":"org",
   "terms":[
      "Organization",
      "subOrganizationOf",
      "transitiveSubOrganizationOf",
      "hasSubOrganization",
      "purpose",
      "classification",
      "identifier",
      "linkedTo",
      "FormalOrganization",
      "OrganizationalUnit",
      "hasUnit",
      "unitOf",
      "location",
      "OrganizationalCollaboration",
      "ChangeEvent",
      "originalOrganization",
      "changedBy",
      "resultedFrom",
      "resultingOrganization"
   ],
   "color":"#3E7EB6",
   "name":"The Organization Ontology",
   "description":"Ontologia utilizada para descrição de
     organizações e estruturas organizacionais,
     incluindo organizações governamentais."
}
```


Figura 26: A estrutura com dados de um vocabulário no formato *JSON*.

A identificação da anotação permite, por exemplo, saber em quais páginas determinado recurso "foi anotado" semanticamente. Por exemplo, em um caso com 100 documentos, no qual 13 deles tiveram o conteúdo anotado sobre um recurso do tipo pessoa. É possível, ao acessar a página do recurso, identificar quais são os 13 documentos em que o recurso foi anotado, além de obter informações sobre os documentos, como título, assunto, URI.

O controle das informações sobre as anotações é realizado pelo componente de gerenciamento de anotações, o qual implementa métodos de geração de identificadores únicos, de edição e exclusão e também de armazenamento das mesmas com metadados, como a data de criação.

O *Web Annotation Data Model* (SANDERSON; CICCARESE; YOUNG, 2015) é utilizado na descrição das anotações. Na ferramenta de anotação semântica, o vocabulário

Open Annotation[3], que representa o modelo de dados de anotações web, é identificado pelo prefixo "*oa*".

Cada anotação gerada pela aplicação é uma instância da classe *oa:Annotation*, possuindo a propriedade *oa:hasTarget* para indicar qual o URI da página em que ela foi gerada e a propriedade *oa:hasBody* para indicar o URI do recurso, que é o sujeito naquela anotação.

```
<http://example.org/Annotation/143942> a oa:Annotation ;
    oa:hasTarget <http://example.org/Article/noticia-1> ;
    oa:hasBody <http://example.org/Organization/Apple> ;
    dct:created ‘‘2015-10-18T00:00:00’’^^xsd:dateTime .
```

Esse conjunto de triplas RDF permite identificar a relação entre determinado recurso com as páginas em que fora anotado. Por exemplo, o recurso `<http://example.org/Organization/Apple>` possui três anotações diferentes, realizadas em três notícias diferentes (*oa:hasTarget*).

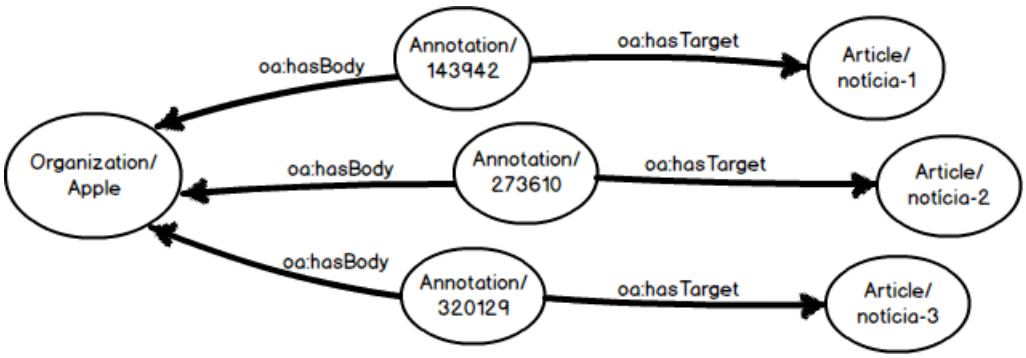


Figura 27: Grafo exibindo um recurso, suas anotações e as páginas onde aparece

O identificador único da anotação é gerado utilizando a função `uniqid()` da linguagem PHP e é atribuído a ela por meio da chamada de um método, utilizando *AJAX*, que trata a requisição de modo assíncrono, enquanto o bloco da anotação permanece ativo na interface gráfica.

No momento da realização da anotação a ferramenta cria um elemento `<span>` envolvendo o conteúdo em questão. Na propriedade *id* deste elemento HTML é atribuído o identificador da anotação, e na propriedade *class* é adicionado o prefixo do vocabulário utilizado para descrever o conteúdo anotado. Isso permite que a região anotada seja identificada com a cor do vocabulário utilizado na mesma.

O conteúdo do bloco de anotação é armazenado temporariamente em um objeto *JSON*, com a seguinte estrutura:

[3] <http://www.w3.org/ns/oa#>

```
{
   "Annotation":{
      "_id":"273610",
      "uri":"http://example.org/Annotation/273610",
      "target":"http://example.org/Article/noticia-2",
      "created":"2015-10-18T00:00:00"
   },
   "Resource":{
      "uri":"http://example.org/Organization/Apple",
      "rdf":[
         {
            "predicate":"dct:title",
            "object":"Apple Inc."
         },
         {
            "predicate":"schema:url",
            "object":"http://apple.com"
         }
      ]
   }
}
```


Figura 28: Conteúdo da anotação estruturado em JSON

Este conteúdo é armazenado no repositório de anotações e também submetido ao componente *Parser* para a que seja feita a conversão do conteúdo em triplas RDF e para, em seguida, armazená-las no banco de dados RDF.

3.2.4 Componente *Parser RDF*

O componente responsável pela conversão dos formatos de dados na aplicação é chamado de *Parser* RDF. Este componente está na camada *Controller*, na arquitetura, e é quem lida com os dados, transformando-os no modelo RDF, antes de serem armazenados no banco de dados RDF.

O *Parser RDF* implementa dois métodos principais para a transformação dos dados: o *annotation2RDF()* e o *resource2RDF()*. Estes métodos processam os dados das anotações realizadas pelo usuário na camada *View* e organizadas na aplicação conforme a estrutura do objeto *JSON* da figura 28.

Os dados referentes à anotação transformados em triplas RDF utilizam um conjunto de termos do vocabulário *Open Annotation* para a descrição. O recurso "anotação" é do tipo *oa:Annotation* e possui as propriedade *oa:hasTarget*, referente à página na qual a anotação foi criada, *oa:hasBody*, aponta para o URI do recurso anotado, e *dct:created*, que possui a data e hora da geração da anotação no padrão ISO 8601.

Os dados sobre o recurso anotado no conteúdo da página são convertidos para RDF

levando em conta o vocabulário, os predicados e objetos definidos pelo usuário que realizou a anotação semântica e não um vocabulário definido na aplicação, como acontece no caso da aplicação, em que se utiliza o *Open Annotation*. O *resource2RDF* processa os dados localizados no atributo "*Resource*" do objeto *JSON* e gera as triplas RDF considerando os valores de *predicate* e *object*. Considerando o exemplo da 28, o método *resource2RDF* retornaria as seguintes triplas RDF:

```
<http://example.org/Organization/Apple> dct:title ‘‘Apple Inc.’’ ;
    schema:url <http://apple.com> .
```

Após a geração das triplas RDF sobre a anotação e sobre o recurso apontado, os dados podem ser armazenados no banco de dados *triplestore*. Para o protótipo da ferramenta sendo desenvolvido durante a pesquisa, estão sendo considerados três diferentes bases de dados, com suporte ao modelo RDF: *Open Link Virtuoso*[4], *AllegroGraph*[5] e o *Neo4J*[6]. As duas primeiras opções possuem suporte nativo ao modelo RDF e ao uso de consultas, na linguagem *SPARQL* e o *Neo4J* é um banco que segue um modelo orientado a grafos, porém sem suporte nativo ao modelo RDF, sendo necessário o uso de bibliotecas de terceiros para trabalhar com RDF e consultas *SPARQL*.

3.2.5 Tipos de dados dos objetos

Uma tripla RDF é composta por sujeito, predicado e objeto. O valor representado como "objeto" da tripla pode ser tanto um identificador IRI (extensão do URI que aceita caracteres internacionalizados, UTF-8) como um *Literal*. Os valores do tipo literal são quaisquer valores que não sejam um IRI, ou seja, podem ser "um texto", "3.14159", "1991-10-27"; no entanto, se o tipo de cada um desses valores não for definido explicitamente, eles podem ser interpretados como uma sequência de caracteres e não como um número decimal ou uma data.

Os tipos de dados definidos para os valores dos objetos em RDF estão disponíveis no *XML Schema 1.1* (PETERSON et al., 2012). Os tipos definidos no *XML Schema* são chamados de *built-in datatypes*[7], e quando utilizados em documentos RDF, os *datatypes* complementam o valor definido na posição "objeto" da tripla, do seguinte modo:

```
‘‘Valor textual’’^^xsd:string
‘‘3.14159’’^^xsd:double
```

[4] <http://virtuoso.openlinksw.com/>
[5] <http://allegrograph.com/allegrograph/>
[6] <http://neo4j.org>
[7] Uma adaptação para português, segundo os autores, seria "tipos de dados primitivos"

O prefixo *xsd* é comumente utilizado para o *namespace XML Schema Datatypes*.

No momento de criação da anotação semântica é possível escolher o tipo de dados para o objeto de cada tripla gerada. Caso não seja especificado nenhum tipo de dado específico, a aplicação tenta identificar se o objeto é um IRI/URI, buscando pelo trecho `http://` no valor, ou então assume-se que o valor digitado é do tipo *xsd:string*.

Na seção 5.1[8] do documento *RDF 1.1 Concepts and Abstract Syntax* (CYGANIAK; WOOD; LANTHALER, 2014) há uma lista com os tipos de dados e suas descrições para serem utilizados com RDF. Essa lista contempla tipos como *xsd:string*, *xsd:double*, *xsd:datetime* e *xsd:boolean*.

3.2.6 Componente para realização da anotação

O elemento de anotação é o que possibilita a interação do usuário com a aplicação, permitindo que ele realize as anotações semânticas sobre determinados conceitos ou partes do conteúdo do documento publicado. Este componente é organizado arquiteturalmente na camada *View* e interage com outros componentes da camada *Controller*.

A ferramenta de anotação funciona com base em um editor de texto, no navegador do tipo *WYSIWYG* (*What You See I What You Get*, que numa adaptação para o português seria "O que você vê é o que você obtém"), desenvolvido a partir do projeto de código aberto *Medium-Editor*. O componente de anotação é acionado quando parte do conteúdo presente no editor é selecionada; desse modo, aparece na tela uma barra de opções.

Além das opções de formatação do texto, como botões para torná-lo negrito ou itálico, é disponibilizado um botão para acionar o bloco de anotação: o bloco de anotação semântica é constituído por três elementos principais: a caixa de texto com a URI do recurso e as outras duas caixas de texto para inserção do predicado e do objeto, para formarem a tripla RDF.

O URI do recurso anotado é obtido através de componente de gerenciamento de URI, quando o usuário define o tipo do recurso anotado:

```
<recurso> rdfs:type <tipo do recurso>
```

. Com a informação sobre o tipo do recurso, o componente responsável pelo gerenciamento de URI gera um identificador para ele, levando em conta o conteúdo sendo anotado e seu tipo. Caso o conteúdo anotado já exista como um recurso na base de dados da aplicação, o usuário pode escolher um identificador já existente para o mesmo.

[8] <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#section-Datatypes>, acessado em 30 de Dezembro de 2015

Após a definição do URI e do tipo (*rdfs:type*), novas triplas podem ser adicionadas para descrever o recurso, bastando apenas preencher os campos "predicado" e "objeto" com os valores correspondentes. Todos os termos dos vocabulários disponíveis no repositório de vocabulários da aplicação podem ser utilizados para complementar a descrição do recurso anotado semanticamente.

A implementação deste componente é realizada com a linguagem, com o uso da *Javascript* e com a utilização da biblioteca *Medium-editor*. O botão e o formulário referentes à ação de anotação são criados como extensões para serem utilizadas com a biblioteca. O código e a documentação do editor *WYSIWYG Medium-editor* estão disponíveis no *Github*, no endereço: <https://github.com/yabwe/medium-editor>.

3.3 Tecnologias consideradas para a implementação

Durante a pesquisa foram estudadas algumas tecnologias e abordagens para serem utilizadas no desenvolvimento da ferramenta de anotação semântica.

Como prova de conceito vem sendo desenvolvido, ao longo da pesquisa, um protótipo da ferramenta proposta no trabalho. Para a implementação dos elementos da camada de visualização desse protótipo utilizou-se a linguagem *Javascript* e o código fonte do editor *Medium-editor*, disponibilizado de modo livre e aberto, além do formato *JSON*, utilizado para a troca de dados entre os componentes da aplicação.

Para os componentes da camada *Controller* - os que fazem parte do *back end*[9] da aplicação - considerou-se a utilização da linguagem PHP e da biblioteca *EasyRDF*, utilizada para produção e consumo de dados no modelo RDF. A biblioteca permite a conversão entre diferentes formatos de serialização de triplas RDF, como *RDF/XML*, *Turtle*, *JSON-LD*, realiza o mapeamento de dados de um grafo RDF em objetos da linguagem PHP, facilitando assim a manipulação dos dados, possui também mecanismos para trabalhar com a persistência dos dados, tratando de consultas *SPARQL* e acessos aos bancos de dados RDF.

O armazenamento dos dados é realizado em três bancos, com propósitos distintos: um banco de dados relacional, um para triplas RDF e um orientado a documentos.

O conteúdo dos artigos e notícias publicados são armazenados em um banco de dados relacional (modelo *RDBMS*), como *MySQL* ou *PostreSQL*. O banco de dados orientado a documentos abriga os repositórios de URI, de anotações e de vocabulários da aplicação e o banco sugerido para implementação é o *MongoDB*, o qual armazena os documentos com a representação em *JSON*. As triplas RDF, geradas a partir das anotações

[9] Geralmente organiza-se uma aplicação cliente-servidor em *front end* e *back end*, representando respectivamente a camada de visualização, visível para o usuário, e as camadas invisíveis que compõe o núcleo do sistema.

semânticas realizadas na aplicação, são armazenadas em um banco de dados específico, chamado de banco de dados RDF ou *triplestore*. Para essa finalidade são considerados dois bancos de dados, com suporte nativo ao modelo RDF e com serviço para realização de consultas *SPARQL*. O *Open Link Virtuoso* e o *AllegroGraph* são as sugestões para uso na ferramenta. Embora sejam comerciais, ou seja, suas licenças de uso devem ser compradas, ambos possuem versões livres, com algumas limitações, como número de triplas, que no caso do *AllegroGraph* é de cinco milhões, na versão não comercial.

3.4 Cenário para implementação de projeto piloto

A implementação de um projeto piloto viabiliza a continuidade do trabalho iniciado nesta pesquisa. Para isso, surge como cenário adequado para a implantação de um sistema para geração de dados estruturados o ambiente de publicação de conteúdos na Web do Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

O NIC.br possui um departamento de comunicação responsável pelas atividades de publicação de conteúdos, nos sites do Comitê Gestor da Internet no Brasil (CGI.br), do próprio NIC.br e das demais áreas da organização (CETIC.br, CERT.br, Ceptro.br, Ceweb.br, Registro.br). Há mais de 10 mil páginas de conteúdo, envolvendo desde artigos, notícias, notas e *releases* para imprensa, publicadas ao longo dos anos, nos sites da instituição.

O contato direto dos autores com a organização possibilita que um projeto piloto voltado para publicação de dados estruturados na Web seja implementado e acompanhado no NIC.br. O desenvolvimento do projeto pode contar com a experiência dos times de Desenvolvimento de Sistemas e do Centro de Estudos sobre Tecnologias Web (Ceweb.br).

Uma das possibilidades a serem exploradas com a ferramenta no cenário do NIC.br seria a geração de dados estruturados a partir do conteúdo das pautas das reuniões do CGI.br podendo, desse modo, ser anotado semanticamente o conteúdo, descrevendo os membros participantes, os temas debatidos, as ações implantadas e os eventos discutidos nos encontros.

Em entrevista transcrita no apêndice A, o coordenador de comunicação da instituição, Everton Teles, vislumbra como possível resultado positivo da implementação da ferramenta, a organização das informações publicadas, contribuindo, inclusive, com as atividades de redação de conteúdo, a partir do conteúdo gerado pela aplicação.

O acompanhamento e desenvolvimento do projeto junto ao NIC.br resultará em novas contribuições para a área pesquisada, levando em consideração a abordagem de simplificação das ferramentas relacionadas com Web Semântica a fim de torná-las mais adequadas, em termos de usabilidade, para os usuários finais.

Considerações finais

A Web Semântica surgiu causando bastante entusiasmo na comunidade, como uma tacada certeira de Tim Berners-Lee, que já criara a Web; porém, sem aparentes resultados imediados, passou a ser questionável.

Os assuntos relacionados com Inteligência Artificial eram emergentes na época em que apareceram as primeiras ideias em torno do que seria proposto e isso, possivelmente, pode ter contribuído para o aumento da expectativa de que a Web Semântica possibilitaria a comunicação autônoma entre máquinas, através da Web, com o propósito de atender às necessidades das pessoas.

Os primeiros longos anos após o surgimento da área foram dedicados, principalmente, para a especificação de padrões. O processo de definição de uma recomendação leva certo tempo, chegando a ser comum a duração do ciclo de criação de um padrão durar cerca de cinco anos. Com a Web Semântica foi desse modo, muitos padrões tiveram que ser especificados para possibilitar a construção de tecnologias que contemplassem as camadas da pilha tecnológica. Houve e ainda há grande esforço por parte das iniciativas relacionadas com Web Semântica para trabalhar em direção à construção e definição de padrões.

Devido ao longo período para a definição dos muitos padrões e as raras aplicações impactantes e relevantes construídas, alguns questionamentos surgiram, em relação ao real resultado proporcionado pela Web Semântica.

Com o levantamento bibliográfico e os estudos sobre o tema, foi possível perceber que se levaram aproximadamente dez anos para que aparecessem mais resultados e aplicações relacionadas com Web Semântica, possivelmente devido ao fim do período de formação dos padrões considerados fundamentais.

Outro ponto interessante observado é que grande parte das publicações datadas entre 2001, ano da publicação do artigo seminal, e aproximadamente 2008 e 2009, que citam o termo Web Semântica, possuem como característica o tempo verbal de algumas orações no futuro, formando frases como “a Web Semântica permitirá”, “com a Web Semântica será possível”, entre outras. Nota-se que publicações mais recentes – dos últimos cinco anos - exibem, cada vez mais, resultados da utilização e aplicação de conceitos e tecnologias relacionados à Web Semântica. Além das publicações e experimentos acadêmicos, há diversas empresas apresentando soluções comerciais, como foi apresentado no estado da arte desta pesquisa.

Observamos que após aproximadamente dez anos de definição e criação de reco-

mendações, houve uma mudança nessa área, passando a existir mais soluções e casos reais, empregando conceitos de Web Semântica.

Abordamos, na pesquisa, Web Semântica mais próxima possível do que fora classificado como Web de Dados. Não tratamos o tema com a visão possivelmente oriunda da IA, de que a Web Semântica possibilitaria a comunicação autônoma entre máquinas, na Web, realizando tarefas e atendendo às necessidades das pessoas. Lidamos com a ideia da publicação de dados estruturados em um ambiente interoperável. Os dados estruturados são descritos por ontologias e vocabulários compartilhados, contribuindo, assim, para a construção da Web de Dados.

Essa abordagem viabiliza a realização de um objetivo antigo que surgiu com a proposta na qual aplicações de software podem atender demandas das pessoas. As aplicações podem fazer uso, consumindo e publicando, do grafo de dados estruturados, com o propósito de facilitar e até mesmo otimizar algumas tarefas, como mostrado em exemplos listados no estado da arte do trabalho.

Este trabalho possui alguns pontos que podem ser aperfeiçoados e explorados em trabalhos futuros, como a possibilidade de realização do processo de anotação semântica, de modo semi-automático, com a identificação de recursos anotados em outros documentos.

O embasamento teórico construído com a pesquisa permite que novos passos sejam dados para aprofundar os estudos nessa área. O processo de construção de conhecimento propiciado pela realização deste trabalho foi extremamente satisfatório e permite a exploração mais profunda, encaminhando para novas etapas de uma contínua pesquisa sobre o assunto.

O objetivo desta pesquisa foi a investigação da viabilidade técnica do desenvolvimento de uma ferramenta de anotação semântica que permita aos usuários publicadores de conteúdo colaborarem com o ecossistema de Web Semântica e conforme a proposta apresentada no capítulo 3, em conjunto com as tecnologias sugeridas para a sua implementação, podemos afirmar que a construção de tal ferramenta se mostra viável; esta constatação, inclusive, é um dos fatores que permitem a continuidade da pesquisa, em trabalhos futuros.

Referências

ADIDA, B. et al. *RDFa 1.1 Primer*. W3C, 2012. Disponível em: <http://www.w3.org/TR/2012/NOTE-rdfa-primer-20120607/>. Citado na página 47.

ADRIAN, B. et al. Epiphany: Adaptable rdfa generation linking the web of documents to the web of data. *Knowledge Engineering and Management by the Masses*, Springer, v. 6317, p. 178–192, 2010. Disponível em: <http://www.springerlink.com/content/c77l6g3nk1727682/>. Citado 4 vezes nas páginas 15, 55, 56 e 57.

ALESSO, H.; SMITH, C. *Thinking on the Web: Berners-Lee, Gödel and Turing*. Wiley, 2006. ISBN 9780471768142. Disponível em: <https://books.google.com.br/books?id=p6CPuAAACAAJ>. Citado 3 vezes nas páginas 11, 13 e 21.

ANDREWS, P.; ZAIHRAYEU, I.; PANE, J. A Classification of Semantic Annotation systems. *Semantic Web Journal*, p. 27, 2011. Disponível em: <http://www.semantic-web-journal.net/sites/default/files/swj123_6.pdf>. Citado 2 vezes nas páginas 22 e 39.

BERNERS-LEE, T. *The original proposal of the WWW, HTMLized*. 1989. Disponível em: <http://www.w3.org/History/1989/proposal.html>. Acesso em: 18 jun. 2015. Citado 2 vezes nas páginas 25 e 26.

BERNERS-LEE, T.; FIELDING, R.; FRYSTYK, H. *RFC 1945, Hypertext Transfer Protocol – HTTP/1.0*. 1996. Disponível em: <http://tools.ietf.org/html/rfc1945>. Citado na página 26.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. *RFC 2396, Uniform Resource Identifiers (URI): Generic Syntax*. 1998. Disponível em: <http://tools.ietf.org/html/rfc2396>. Citado na página 30.

BERNERS-LEE, T.; FISCHETTI, M. *Weaving the web: The past, present and future of the World Wide Web by its inventor*. [S.l.: s.n.], 2000. Citado 2 vezes nas páginas 27 e 29.

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001. Citado 5 vezes nas páginas 11, 13, 21, 22 e 28.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, p. 205–227, 2009. Citado na página 29.

CASTELLS, M. *A Galáxia Internet: reflexões sobre a Internet, negócios e a sociedade*. [S.l.]: Zahar, 2003. Citado na página 27.

CERN. *Tim Berners-Lee's proposal*. 2008. Disponível em: <http://info.cern.ch/Proposal.html>. Acesso em: 18 jun. 2015. Citado 2 vezes nas páginas 15 e 25.

CYGANIAK, R.; WOOD, D.; LANTHALER, M. *RDF 1.1 Concepts and Abstract Syntax*. 2014. Disponível em: <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#section-Datatypes>. Acesso em: 30 dez. 2015. Citado na página 73.

FIELDING, R. et al. *RFC 2616, Hypertext Transfer Protocol – HTTP/1.1*. 1999. Disponível em: <http://tools.ietf.org/html/rfc2616>. Citado na página 26.

GARCIA, P. S. R. *Estudo sobre aplicação de web semântica e visualização em dados abertos*. Dissertação (Masters) — PUC-SP, 2011. Citado na página 27.

GIL, A. C. *Como elaborar projetos de pesquisa - 4. ed.* São Paulo: Editora Atlas S.A., 2002. ISBN 85-224-3169-8. Citado 2 vezes nas páginas 13 e 21.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993. Citado 2 vezes nas páginas 22 e 36.

GUIZZARDI, G. *Ontological foundations for structural conceptual models*. [S.l.]: CTIT, Centre for Telematics and Information Technology, 2005. Citado 4 vezes nas páginas 11, 13, 22 e 36.

HARRIS, S.; SEABORNE, A. *SPARQL 1.1 Query Language*. 2013. Disponível em: <http://www.w3.org/TR/sparql11-query/>. Citado na página 32.

HITZLER, P. et al. *OWL 2 Web Ontology Language Primer (Second Edition)*. 2012. Disponível em: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>. Citado 2 vezes nas páginas 37 e 38.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados*. [S.l.]: Novatec Editora, 2015. Citado 6 vezes nas páginas 11, 13, 22, 28, 36 e 37.

KARGER, D. The semantic web and end users: What's wrong and how to fix it. *Internet Computing, IEEE*, v. 18, n. 6, p. 64–70, Nov 2014. ISSN 1089-7801. Citado 4 vezes nas páginas 21, 22, 23 e 59.

KUROSE, J.; ROSS, K. *Redes de computadores e a internet: uma abordagem top-down*. [S.l.]: Pearson, 2010. ISBN 9788588639973. Citado 2 vezes nas páginas 26 e 27.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. *Data on the Web Best Practices*. 2015. Disponível em: <http://www.w3.org/TR/2015/WD-dwbp-20151217/>. Acesso em: 2 jan. 2016. Citado na página 38.

M., D.; M., S. *RFC 3987, Internationalized Resource Identifiers (IRIs)*. 2005. Disponível em: <http://tools.ietf.org/html/rfc3987>. Citado na página 30.

MANOLA, F.; MILLER, E.; MCBRIDE, B. *RDF Primer*. 2004. Disponível em: <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#intro>. Citado 4 vezes nas páginas 15, 22, 31 e 32.

MASOLO, C. et al. Ontology library (wonder-web deliverable d18). *URL: http://www. loa-cnr. it/Papers D*, v. 18, 2003. Citado na página 36.

MCCUINNESS, D. L. Ontologies come of age. *Spinning the semantic web: bringing the World Wide Web to its full potential*, 2003. Citado na página 37.

NELSON, T. H. Getting it out of our system. *Information retrieval: A critical review*, Thomson Books, Washington DC, v. 191, p. 210, 1967. Citado na página 26.

PEIRCE, C. S.; HARTSHORNE, C.; WEISS, P. Collected papers of charles sanders peirce: Vol. iii, exact logic. 1935. Citado na página 36.

PENA, R. A. P.; SCHWABE, D. *Suporte semântico à publicação de conteúdo jornalístico na Web*. Dissertação (Mestrado) — PUC-Rio, 2012. Citado 6 vezes nas páginas 15, 49, 50, 51, 52 e 53.

PETERSON, D. et al. *W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes*. 2012. Disponível em: <http://www.w3.org/TR/xmlschema11-2/>. Acesso em: 30 dez. 2015. Citado na página 72.

RAGGETT, D. et al. *Raggett on HTML 4 (2nd Ed.)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1998. ISBN 0-201-17805-2. Disponível em: <http://www.w3.org/People/Raggett/book4/ch02.html>. Citado na página 27.

RECORDON, D. 2010. Disponível em: <https://www.scribd.com/doc/30715288/The-Open-Graph-Protocol-Design-Decisions>. Acesso em: 17 out. 2015. Citado na página 47.

RODRÍGUEZ-ROCHA, O. et al. Semantic annotation and classification in practice. *IT PROFESSIONAL*, IEEE, v. 17, n. IT-Ena, p. 33–39, 2015. Disponível em: <http://porto.polito.it/2585561/>. Citado 6 vezes nas páginas 11, 13, 21, 22, 28 e 38.

SÁNCHEZ, D.; ISERN, D.; MILLAN, M. Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, Springer-Verlag, v. 27, n. 3, p. 393–418, 2011. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/s10115-010-0302-3>. Citado na página 38.

SANDERSON, R.; CICCARESE, P.; YOUNG, B. *Web Annotation Data Model*. 2015. Disponível em: <http://www.w3.org/TR/annotation-model/>. Acesso em: 29 dez. 2015. Citado na página 69.

SANTAELLA, L.; VIEIRA, J. A. *Metaciência como guia da pesquisa: uma proposta semiótica e sistêmica*. Rio de Janeiro: Mérito Editora, 2008. ISBN 9788561758257. Citado 3 vezes nas páginas 11, 13 e 21.

SANTOS, F. C.; CARVALHO, C. L. de. *Aplicações de Suporte à Web Semântica*. [S.l.], 2007. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-07.pdf>. Citado na página 37.

SHADBOLT, N.; HALL, W.; BERNERS-LEE, T. The semantic web revisited. *Intelligent Systems, IEEE*, IEEE, v. 21, n. 3, p. 96–101, 2006. Citado na página 36.

SILVA, V. M. da. Revisão sistemática da evolução mvc na base acm. 2012. Disponível em: <https://www.researchgate.net/profile/Valeria_Silva8/publication/264003410_Reviso_sistemtica_da_evoluo_MVC_na_base_ACM/links/00b4953c839fa9bdd2000000.pdf>. Citado na página 64.

SLIMANI, T. Semantic annotation: The mainstay of semantic web. *CoRR*, abs/1312.4794, 2013. Disponível em: <http://arxiv.org/abs/1312.4794>. Citado 3 vezes nas páginas 21, 22 e 39.

SMITH, B.; WELTY, C. Ontology: Towards a new synthesis. In: ACM PRESS, USA, PP. III-X. *Formal Ontology in Information Systems*. [S.l.], 2001. p. 3–9. Citado na página 36.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a ciência da informação. *Ciência da Informação, Brasília*, SciELO Brasil, v. 33, n. 1, p. 132–141, 2004. Citado na página 26.

SPORNY, M.; KELLOGG, G.; LANTHALER, M. *JSON-LD 1.0*. 2013. Disponível em: <http://www.w3.org/TR/2013/WD-json-ld-20130411/>. Citado na página 30.

THOMSON Reuters Open Calais - API User Guide. Thomson Reuters, 2015. Disponível em: <http://www.opencalais.com/wp-content/uploads/2015/10/ThomsonReutersOpenCalaisAPIUserGuide291015.pdf>. Acesso em: 04 nov. 2015. Citado na página 53.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 4, n. 1, p. 14 – 28, 2006. ISSN 1570-8268. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1570826805000338>. Citado 2 vezes nas páginas 22 e 38.

APÊNDICE A – Entrevista

Entrevista realizada com Everton Teles, coordenador de comunicação do Núcleo de Informação e Coordenação do Ponto BR.

Interpreta-se “E.” como fala do entrevistado e “P.” como fala do pesquisador.

P.: Oi Everton, obrigado por ter aceitado este convite para a entrevista. Eu faço pesquisa de mestrado no programa Tecnologias de Inteligência e Design Digital, da PUC-SP, sob a orientação do Professor Demi Getschko. Meu trabalho é sobre Web Semântica.

Minha pesquisa é a respeito do tópico de anotação semântica na Web. Nele faço a proposta de uma ferramenta de anotação semântica para geração de dados estruturados com o uso de ontologias e vocabulários compartilhados.

Um dos fatores que motivam esta minha pesquisa é a possibilidade de inserir mais usuários no ecossistema de Web Semântica, de modo que jornalistas e publicadores de conteúdo passem a publicar metadados na Web do mesmo modo que publicam conteúdos human-readable.

Antes de explicar em mais detalhes a ferramenta proposta, eu gostaria de saber sobre seu contexto, formação, se trabalha com a publicação de conteúdo na Web.

E.: Sou Jornalista, formado pela Universidade Metodista de São Paulo, e tenho experiência de aproximadamente 10 anos, boa parte deste tempo aqui no NIC.br.

Sou coordenador de comunicação do NIC.br, e entre minhas atribuições também estão atender ao Comitê Gestor da Internet no Brasil. O conhecimento que possuo sobre Web Semântica não é profundo e também não é técnico. Pelo fato de trabalhar próximo do W3C Brasil e do Ceweb, passo a ter mais contato com o tema.

Na minha rotina a coordenação e a produção de conteúdo são as tarefas principais na minha rotina. Além disso mantenho artigos institucionais nos canais de redes sociais.

Não realizo, nesse meu processo de publicação de conteúdo, nenhuma tarefa no sentido de tornar aquela informação estruturada. Faço então a produção de conteúdo human-readable.

P.: As informações publicadas pela equipe de comunicação do NIC.br são muito relevantes, principalmente considerando o contexto de trabalho da instituição. Mas esses dados são apenas passíveis de interpretação por humanos, ou seja, human-readable, e não são publicados em formato estruturado.

E.: Passar a produzir e publicar o conteúdo de modo adequado à Web Semântica ainda é um passo que não foi dado aqui no NIC.br.

Produzimos ainda na base Web de Documentos. Publicamos documentos na Web.

O Cetic.br, um departamento do NIC.br, está conduzindo uma inciativa de publicação de dados estruturados e em formato aberto. Porém são dados estatísticos, microdados, e não conteúdo textual. As análises dos indicadores publicados são, geralmente, em PDF.

P.: Vou te apresentar as imagens e esquemas do protótipo da ferramenta de anotação semântica que venho pesquisando no meu trabalho.

Nestas imagens eu mostro o trecho de uma notícia publicada no site do NIC.br. No conteúdo da notícia há informações sobre algumas pessoas e organizações ligadas, de certo modo, com o Núcleo de Informação e Coordenação do Ponto BR.

É possível identificarmos quais nomes e conceitos no texto referem-se a pessoas, organizações, endereços, etc. Porém para máquina isso não passa de um conjunto de caracteres, que podem ser interpretáveis, porém não há um contexto descrevendo o significado destas strings, as quais nós, humanos, conseguimos identificar e associar alguma semântica.

Como você sabe, um elemento fundamental para a Web Semântica é o modelo RDF. Um modelo para intercâmbio de dados em formato de triplas sujeito, predicado e objeto.

Neste exemplo, conseguimos identificar no texto o nome da organização *Berkman Center*, mas para a máquina, é apenas um conjunto de caracteres. A ferramenta permite associar uma ontologia para descrever o significado desse conjunto de caracteres. Nesse caso estou utilizando a ontologia do W3C chamada *The Organization Ontology*. O resultado dessa associação é uma, ou mais, triplas RDF, assim, o conteúdo passa a ser machine-readable.

Após um grande conjunto de dados serem publicados em formato machine-readable, é possível exibir tais dados estruturados em diferentes aplicações

Eu gostaria de saber suas impressões sobre este exemplo que te mostrei.

E.: Ótimo. Antes de apresentar minhas impressões, gostaria de fazer uma pequena observação sobre a área de jornalismo.

Temos um número expressivo de jornalistas sendo formados todo ano. Com uma oferta muito grande de jornalistas, pode ser que não tenhamos "ouvidos suficiente para prestar atenção no que todos estão falando", desde distinguir sinais de ruído.

Então, muito conteúdo é publicado, mas nem sempre é possível dar conta da interpretação desse conteúdo, muitas vezes relevante, publicados.

Onde vejo com bons olhos o uso de uma ferramenta deste tipo é numa demanda

que parece ser necessária para todos que publicam conteúdo na Web, no sentido de auxiliar na redação de um conteúdo, permitindo filtrar e catalogar numa base todos os conteúdos relacionados relacionados ao tema escrito. Algo nesse sentido, se for possível desenvolver com Web Semântica, poderia ser muito útil.

P.: Entendi essa sua observação, talvez esteja até mais relacionada com o consumo dos dados estruturados, com o desenvolvimento de aplicações que tratem dessa parte e isso é algo que pretendo tratar na continuidade desta pesquisa.

Um dos pontos que investigo ao conduzir minha pesquisa nesta área é a possibilidade de aproximação de ferramentas de Web Semântica dos usuários finais.

Como no exemplo que te mostrei, não é necessário ter conhecimentos específicos, como de linguagens XML, JSON-LD e outras, para gerar conteúdo em RDF. Isso permite que usuários finais, nesse caso, jornalistas e redatores de conteúdo, publiquem dados estruturados sem necessariamente ter conhecimentos específicos de linguagens e tecnologias de Web Semântica.

E.: Mas esse conteúdo anotado pode ser de qualquer página? Se eu acessar a página de um portal de notícias posso anotar semanticamente as entidades que eu eventualmente encontrar no texto?

P.: Não. A ferramenta é um editor de texto que permite que você anote o conteúdo que você está gerando, podendo esse ser de sua autoria ou de outra fonte, mas deve necessariamente ser "editado"pela ferramenta, somente assim é possível anotar.

Além disso, o conteúdo anotado semanticamente alimenta uma base de dados própria, definida para seu escopo. E não numa base compartilhada publicamente na Web.

E.: Mas esse conteúdo anotado não fica disponível na Web?

P.: Sim, o conteúdo anotado também fica disponível na Web, distribuído tanto por HTML, a versão human-readable, quando em JSON-LD ou RDFa a versão machine-readable.

As bases de dados específicas, definidas para um escopo ou projeto, servem, além de disponibilizar os dados machine-readable, também para criar aplicações como grafos de recomendações e outras.

E.: Entendido. Se eu atribuo um vocabulário numa palavra em um primeiro momento, nas próximas vezes que essa palavra aparecer eu devo atribuir o vocabulário novamente ou ele reconhece automaticamente?

P.: Nesta proposta da ferramenta não acontece a anotação automaticamente. Porém, como continuidade da pesquisa o trabalho focará na implementação de um método para anotação semi-automática, conforme os três tipos de anotações que apresentei na pesquisa, manual, semi-automática e automática.

E.: Interessante.

Do ponto de vista de produção de conteúdo e também no meu papel de coordenação de produção de conteúdo, onde devo acompanhar o que tem sido produzido, ver como estão estruturadas as informações nos textos, sobre quais pessoas foram publicados os últimos artigos, quais temas são mais frequentes e uma infinidade de métricas e relatórios.

Imagino que com uma possível ferramenta desse tipo, seria possível gerar relatórios de modo automatizado contendo informações como as que te falei, por exemplo em quais contextos ou assuntos determinada pessoa foi citada nos últimos três meses.

P.: Com os dados estruturados publicados e armazenados adequadamente passa a ser possível criar aplicações que faça o consumo desses dados a fim de gerar relatórios como esses que você mencionou.

Além disso, um exemplo que utilizei e também um possível cenário que especulamos durante a pesquisa foi o seguinte: a anotação semântica do conteúdo das atas das reuniões do Comitê Gestor da Internet no Brasil.

Vamos, por exemplo, anotar semanticamente o nome de determinado conselheiro em diferentes atas. Esses dados estruturados me permitem gerar uma página específica para este recurso anotado, no caso, o conselheiro e também, a partir do grafo gerado, sei em quais reuniões ele participou, poderia saber também os temas das diferentes reuniões transcritas nas atas. Poderia também adicionar no grafo quais os temas tal conselheiro sugeriu, sobre qual projeto foi discutido.

Então, anotando semanticamente com os vocabulários específicos todo esse conteúdo, nesse caso, produzido pelo NIC.br, seria extremamente útil, podendo gerar muitas formas de consumo e visualizações a partir dos dados estruturados publicados.

E.: Realmente. E isso é um recurso fantástico, tanto para o profissional de comunicação, pois poderia possivelmente contribuir no processo de obtenção de dados para a redação de conteúdos, mas também pode possibilitar um processo bastante transparente, permitindo ao público ver se nós, como assessoria de comunicação para as ações do CGI.br, estamos falando mais sobre determinado assunto do que outro, se estamos publicando mais sobre determinado setor e nada de outro, etc.

Veja essa como uma aplicação bem interessante desta ferramenta, pois pode contribuir não somente no processo de produção de conteúdo, mas também nesse processo de transparência, compilando dados e métricas como as que mencionei.

Porém imagino que deveria ser alterado o fluxo de produção de conteúdo, pois atualmente não realizamos desse modo, pensando em tornar os dados machine-readable.

P.: No estado da arte da minha pesquisa apresentei um caso de uso da Globo.com, no qual foi desenvolvida uma aplicação de Web Semântica e Ontologias para auxiliar na

redação de conteúdo do Globoesporte.com

O objetivo principal desta ferramenta é encurtar os passos para a geração de dados estruturados, ou de triplas RDF.

Pois para você, como jornalista e produtor de conteúdo, não seria necessário utilizar um editor de código e escrever código em uma linguagem específica, como Turtle, para criar as triplas RDF para descrever seu conteúdo.

Esse processo de geração de triplas RDF é auxiliado pela ferramenta proposta na pesquisa.

Everton, minha próxima questão é em relação ao fluxo de trabalho para publicação de conteúdo: como ocorre, vocês redigem o conteúdo em uma aplicação no próprio computador, posteriormente publicam no site?

E.: O processo que nossa equipe adota parte de uma ideia ou discussão inicial para então iniciar a escrita utilizando, geralmente, o Microsoft Word, fazendo as devidas correções e posteriormente quando finalizado, esse conteúdo é enviado para aprovações do coordenador e do gerente de comunicação e também para aprovação da área que demandou a escrita deste conteúdo.

Após todas aprovações e possível alterações no texto, o conteúdo é adicionado aos nossos sites por meio do gerenciador de conteúdo que utilizamos, o qual permite que seja colado o conteúdo no bloco de texto e também que seja personalizado com tags HTML. O conteúdo publicado é armazenado no formato HTML em nossa base de dados e disponibilizado nos nossos sites.

P.: Esta ferramenta que venho estudando na pesquisa pode ser acoplada ao gerenciador de conteúdo que vocês utilizando, o qual permite a edição do conteúdo em HTML.

Everton, eu agradeço pela conversa e muito obrigado por expressar sua opinião e dar sugestões para a minha pesquisa.