UNIVERSIDADE DO ESTADO DA BAHIA
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
COLEGIADO DE SISTEMAS DE INFORMAÇÃO

UMA SOLUÇÃO COMPUTACIONAL PARA INTEGRAÇÃO ENTRE WEB SEMÂNTICA E REDES COMPLEXAS

Shankar Cabus de Teive e Argollo

Salvador
2012

UNIVERSIDADE DO ESTADO DA BAHIA
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
COLEGIADO DE SISTEMAS DE INFORMAÇÃO

UMA SOLUÇÃO COMPUTACIONAL PARA INTEGRAÇÃO ENTRE WEB SEMÂNTICA E REDES COMPLEXAS

Shankar Cabus de Teive e Argollo

Monografia apresentada à Banca Examinadora como exigência parcial para obtenção do título de bacharel em Sistemas Informação pela Universidade do Estado da Bahia.

Orientador:
Eduardo Manuel de Freitas Jorge

Salvador
2012

SHANKAR CABUS DE TEIVE E ARGOLLO

Uma solução computacional para integração entre Web Semântica e Redes Complexas

Monografia apresentada à Banca Examinadora como exigência parcial para obtenção do título de bacharel em Sistemas de Informação pela Universidade do Estado da Bahia.
Orientador. Eduardo Manuel de Freitas Jorge

BANCA EXAMINADORA

Prof. Eduardo Manuel de Freitas Jorge
Doutor em Difusão do Conhecimento
Universidade do Estado da Bahia

Prof. Alexandre Rafael Lenz
Doutorando em Ciência da Computação
Universidade do Estado da Bahia

Prof. Uedson Santos Reis
Mestre em Modelagem Computacional
Serviço Nacional de Aprendizagem Industrial

Agradecimentos

Lista de Figuras

Lista de Algoritmos

Lista de Tabelas

Sumário

1 Introdução

Desde o surgimento da primeira rede de computadores em 1960, a internet vem sofrendo transformações no intuito de facilitar a produção, compartilhamento e organização da informação. Novos conceitos de design e usabilidade, novas ferramentas de navegação e publicação, facilidade de acesso, velocidade na transmissão de dados, são características da Web atual, que têm o objetivo de tornar o uso da internet cada vez mais prazeroso e produtivo.

Com a facilidade para compartilhar e produzir conteúdo, o volume de informação na Web cresce diariamente numa escala cada vez maior. Logo, o desafio da Web é facilitar a recuperação dessas informações. Existem ferramentas de busca como Google, Bing e Yahoo! que se propõe a resolver este problema. Seu funcionamento básico consiste em navegar entre links da web, indexar o conteúdo das páginas visitadas e, de acordo com uma busca feita por um usuário, retornar as páginas mais relevantes. Mas a máquina não consegue interpretar a consulta, apenas aplicar regras em cima de um conjunto qualquer de palavras desconexas. Por isso, buscas mais complexas, que envolvem muitos cruzamentos de dados, diminuem consideravelmente a eficácia desse método de recuperação de dados.

Pensando no problema da organização dos dados, Tim Berners-Lee apresenta, em 2001, o termo Web Semântica, que tem como objetivo transferir a responsabilidade de interpretação da informação, do usuário para a máquina (BREITMAN, 2005). Berners-Lee propõe a estruturação das informações de forma semântica, utilizando metadados e modelos semânticos para interligá-las. A essa nova forma de publicar e interligar os dados, Berners-Lee deu o nome de *Linked Dat*a, partindo da ideia de que o valor e uso dos dados crescem quanto mais eles estiverem interligados (BIZER et al, 2008). Essa nova estrutura possibilita que a máquina interprete os dados, facilitando a organização e consequentemente, a recuperação das informações.

Em Janeiro de 2007 surge o projeto Linking Open Data, com o objetivo de desenvolver *Linked Data* a partir da identificação de dados existentes sob licença livre e convertê-los para o modelo de dados da Web Semântica, o RDF (*Resource Description Framework*). O projeto que começou de forma lenta, vem crescendo a cada ano, passando de 12 conjuntos de dados

em 2007 para 295 em Setembro de 2011. As figuras 1 e 2 mostram a representação gráfica do projeto em 2007 e em 2011, respectivamente.

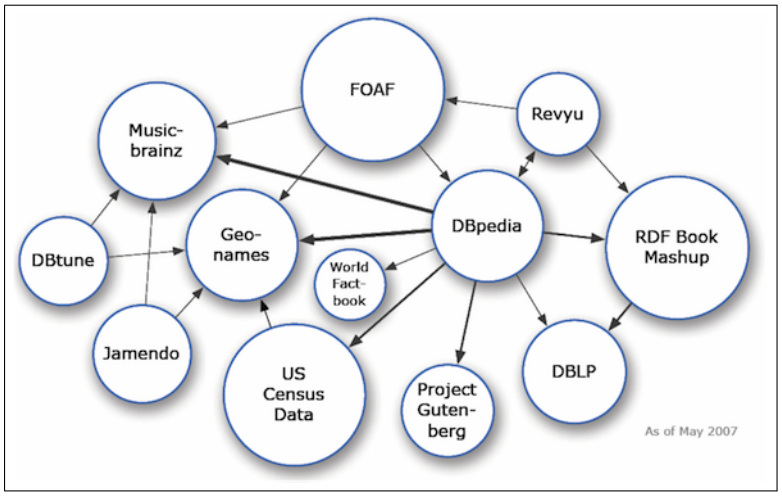

Figura 1: Diagrama *Linking Open Data* em Maio de 2007

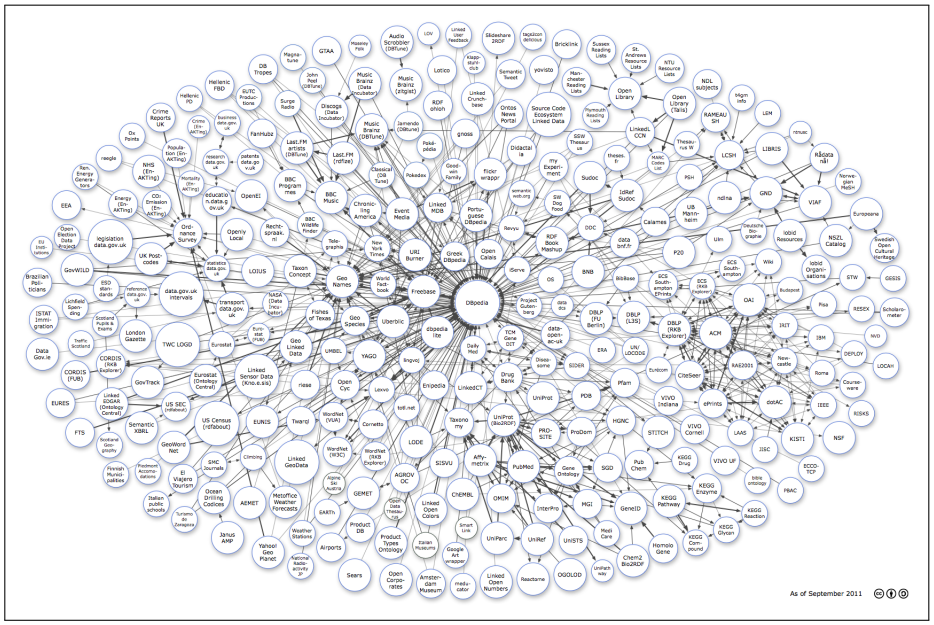

Figura 2: Diagrama *Linking Open Data* em Setembro de 2011

As relações entre os dados do Linked Data podem ser representadas por meio de redes complexas, que consistem em um conjunto de vértices ou nodos ligados por linhas chamadas arestas. Através de ferramentas de análise de redes, é possível fazer uma interpretação gráfica e matemática de propriedades que caracterizam as redes complexas. Contudo, até o presente momento não foram encontradas ferramentas ou protocolos que auxiliem na conversão dos dados no formato RDF para uma entrada legível por estas ferramentas de análise de redes.

Desta forma, este projeto tem por objetivo propor um modelo computacional, chamado de RDFree, para extrair os dados em RDF de bases de dados Linked Data e convertê-los para um formato possível de ser interpretado por ferramentas de análise de redes complexas, integrando os dois campos de pesquisa supracitados: Web Semântica e Redes complexas.

Espera-se alcançar tal objetivo através das seguintes metas ou objetivos específicos:

- Criar uma arquitetura genérica com os elementos de toda a solução do modelo computacional proposto.
- Criar um componente para leitura de dados no padrão RDF, possibilitando a definição de parâmetros que permitam a filtragem e indicação de elementos que são vértices e arestas numa rede semântica.
- Gerar um arquivo de texto num padrão interpretável por softwares de análise de redes semântica.

O trabalho será validado pelo uso do modelo computacional proposto em um determinado conjunto de dados RDF, convertendo os dados para um arquivo de entrada de softwares de análise de redes e, por fim, gerar métricas que caracterizam e definem a topologia da rede utilizada.

Diferentes campos de estudos científicos e sociais serão beneficiados com este trabalho que se propõe a auxiliar na coleta, processamento e análise de dados espalhados pela Web Semântica, provendo subsídios para que pesquisadores possam interpretar os dados e utilizá-los em duas pesquisas.

2 *Fundamentação Teórica*

Este capítulo tem por objetivo apresentar os principais temas deste trabalho. Mas para um melhor entendimento, inicialmente é feita uma contextualização história desde o surgimento da Web até o presente momento em que este trabalho foi escrito. Em seguida, é abordada a Web Semântica e as tecnologias que a rege: RDF e SPARQL. Por fim, apresenta-se os conceitos de Linked Data e redes complexas.

2.1 A História da Web

Segundo (CONNOLLY, 2000), em 1989, Tim Berners-Lee, funcionário do CERN[1] escreveu uma proposta[2] de gerenciamento de informação. Esta proposta se tratava de um grande banco de dados com hiperligações, que mais tarde recebera o nome de World Wide Web (grande teia mundial), também conhecida como WWW ou simplesmente Web.

No final de 1990, Berners-Lee havia colocado o projeto em prática e construído as ferramentas necessárias para o funcionamento da Web: Protocolo de Transferência de Hipertexto (HTTP), responsável pela transferência de dados pela Web; a Linguagem de Marcação de Hipertextos (HTML), utilizada para representar os textos e os hiperlinks das páginas da Web; e ainda o primeiro navegador, servidor HTTP e servidor Web. No ano seguinte, Berners-Lee publicou um resumo do projeto WWW, marcando este ano como o nascimento da Web como um serviço público da Internet.

> O projeto World Wide Web (WWW) tem por objetivo permitir que todas as ligações possam ser feitas com qualquer informação, não importando onde elas se encontrem. [...] O projecto WWW foi lançado para permitir que os físicos de altas energias possam trocar informações, notícias e documentos. Estamos muito interessados em alargar a web a outras áreas e ter servidores de portas de ligação

[1] Conseil Européen pour la Recherche Nucléaire (Organização Europeia para a Pesquisa Nuclear)
[2] http://www.w3.org/History/1989/proposal.html

(Gateway) para outros dados. (BERNERS-LEE, 1991)

Na Web, os recursos, como páginas, documentos e imagens, são identificados por meio de URIs, acrônimo em inglês para *Uniform Resource Identifier* (Identificador Uniforme de Recursos). Sendo que o grande diferencial do projeto de Berners-Lee foi a forma como as páginas eram relacionadas, utilizando um tipo específico de URI para se relacionar com outros recursos, o Uniform Resource Locator (URL). O URL é o endereço de um recurso disponível numa rede, podendo ser a Internet ou até uma rede interna (intranet), e que possui a seguinte estrutura: `protocolo://maquina/caminho/recurso`.

A adoção da World Wide Web por universidades e laboratórios de física, fez com que a Web ganhasse popularidade, como pode ser visto na figura 3, que mostra a quantidade de clientes da Web de julho de 1991 à julho de 1994.

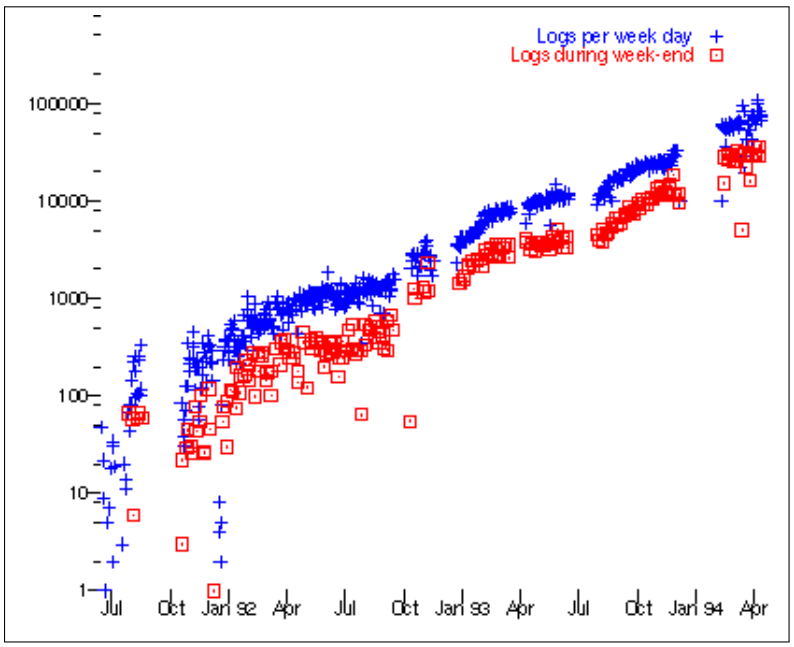

Figura 3: Gráfico de crescimento da quantidade de clientes da Web entre 1991 e 1994. Fonte: (CONNOLLY, 2000)

Entre 1999 e 2001, a utilização da Web continuou crescendo, mas agora com uma motivação ainda maior, a Web começou a ser explorada de forma comercial e marcada por grandes portais geradores de conteúdo. Nesta fase, também conhecida como Web 1.0, os usuários tinham características *read-only* (somente leitura), ou seja, eles utilizavam a Web de forma passiva, consumindo informações geradas e disponibilizadas por grandes sítios que dominavam o mercado Web (BLATTMANN; SILVA, 2007), tais como Altavista, Geocities, Yahoo, UOL, entre outros. Mais tarde, (BREITMAN, 2005) caracterizou esse período como Web Sintática, por

não haver qualquer tipo de interpretação da informação presente na Web, pois o objetivo era somente o de apresentar o conteúdo para o usuário.

Além de informações como notícias, artigos e imagens, outras mídias começaram a surgir e a Web tornou-se também um meio de entretenimento, cercado de jogos, bate-papo, vídeos, entre outros. Dá-se inicio à popularização da Web, que traz consigo uma nova preocupação: dar ao usuário a melhor experiência possível na frente do computador. Então as tecnologias avançaram e surgiram ainda novas linguagens, navegadores, protocolos e tecnologias, como AJAX[3], técnicas de usabilidade foram estudadas e aplicadas, as conexões de internet ficaram cada vez mais rápidas, os sítios cada vez mais atrativos e o usuário participativo. Começa a segunda grande mudança de paradigma da Web, denominada por (., .) como Web 2.0 e também conhecida como Web colaborativa, pois o usuário deixa de ser apenas consumidor de informação, para se tornar um produtor de conteúdo. A interação agora faz parte da Web. É neste período que surgem os blogs (sítios cuja estrutura permite atualização rápida pelo usuário a partir de artigos), e sítios como Youtube, Facebook, Wikipedia, entre outros.

> É a passagem de uma rede estática para uma dinâmica e de colaboraçãp, onde administram grupos de trabalhos colaborativos e interdisciplinares. E? constituída de um conjunto de novos serviços e ferramentas baseadas em um enfoque de colaboração, de dinamismo, e de facilidade de transferência de informação em todos os suportes (texto, imagem, áudio e vídeo). Web 2.0 descreve os sites de serviços ou tecnologias que promovem a capacidade de compartilhar e colaborar na rede. (BORCHANI, 2007).

Apesar do crescimento da Web e mudança de paradigma no que diz respeito à participação do usuário, uma coisa não mudou, a Web continua feita para ser entendida apenas pelo usuário. O que antes não parecia ser problema, agora começa a ficar preocupante, pois, com a facilidade para produzir conteúdo, o volume de informação cresce num ritmo ainda mais acelerado e desorganizado, dificultando a recuperação dessas informações espalhadas entre milhões de sítios. A figura 4 mostra a quantidade de usuários da Internet em milhões no período de 1995 à 2012.

[3] Acrônimo em língua inglesa de *Asynchronous Javascript and XML*, trata-se do uso metodológico de tecnologias como Javascript e XML, providas por navegadores, para tornar páginas Web mais interativas com o usuário, utilizando-se de solicitações assíncronas de dados.

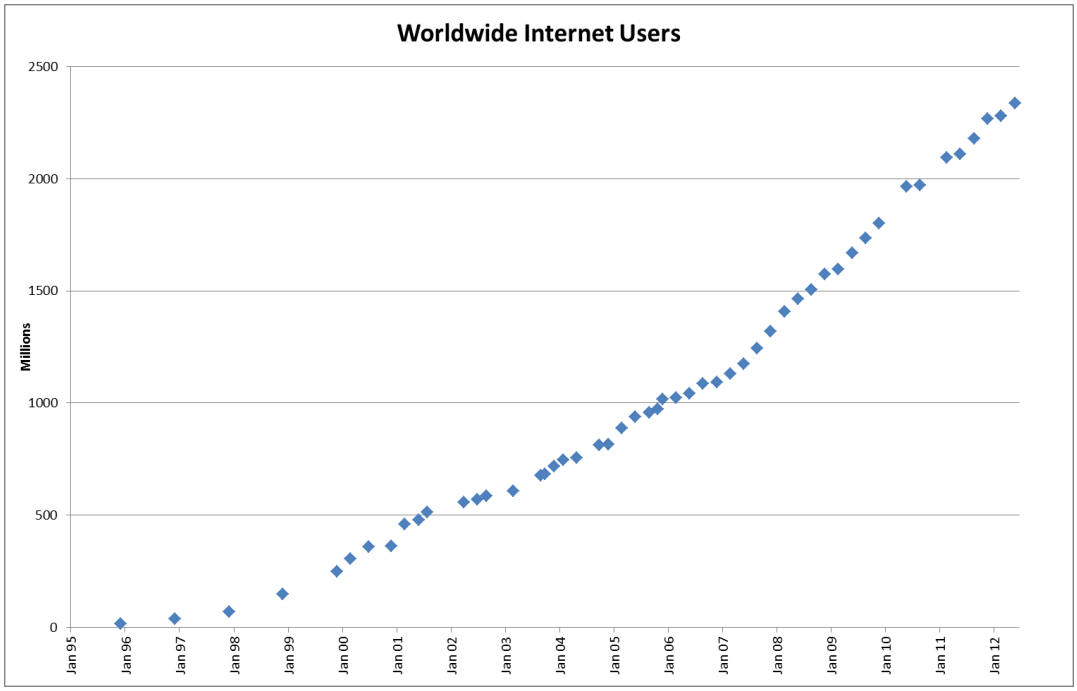

Figura 4: Gráfico de crescimento de usuários da Web entre janeiro de 1995 e janeiro de 2012. Fonte: Internet World Stats

Como as informações estavam espalhadas entre milhões de sites pela Web, era necessário uma forma de encontrar informações específicas. A primeira alternativa foi a criação de portais que agrupassem outros sites e os organizassem por temas, como era o caso do GeoCities, muito popular na década de 1990, porém extinto em 2009, pois já existia um método mais eficaz para recuperar as informações: as ferramentas de busca, também conhecidas como motores de busca. Estas ferramentas, a exemplo do Google, Bing e Yahoo!, percorrem as hiperligações da Web criando um índice dos documentos, que contém, entre outras informações, a quantidade de vezes que cada palavra aparece naquele documento. Dado uma palavra-chave buscada por um usuário, a ferramenta retorna resultados baseados, principalmente, em métricas como densidade e quantidade de vezes que a palavra-chave aparece nos documentos. Contudo, em buscas mais complexas, que envolvem muitos termos ou muitos cruzamentos de dados, esses buscadores convencionais tem sua eficácia reduzida, pois eles não conseguem interpretar o sentido de uma frase ou significado de cada palavra dentro de um contexto.

É pensando no problema da organização e recuperação da informação que em 2001 Berners-Lee, juntamente com James Hendler e Ora Lassila, publica um artigo na revista *Scientific American*, intitulado "Web Semântica: um novo formato de conteúdo para a Web que tem significado para computadores vai iniciar uma revolução de novas possibilidades". Sua proposta é fazer com que as máquinas tenham capacidade de entender o que o usuário deseja por meio de uma estruturação dos dados, que será explicada na próxima seção sobre Web Semântica, também

conhecida como Web 3.0.

2.2 Web Semântica

"Quais bandas inglesas de *rock and roll*, formadas nos anos 80 e que já tocaram no *Rock In Rio*?". Este é um tipo busca que demanda uma compreensão semântica da frase. Um motor de busca precisaria entender que a busca se refere a grupos musicais oriundos da Inglaterra, que por sua vez é um país. Também precisaria compreender que "anos 80"é um período entre 1980 e 1989, que *rock and roll* é um estilo musical e *Rock In Rio* é um evento de *rock and roll*. Além de compreender o sentido da frase, o buscador precisaria cruzar todas essas informações a fim de encontrar o resultado esperado pelo usuário. Este é um tipo de tarefa que as ferramentas de busca encontradas hoje não são capazes de fazer, pois elas se limitam a uma análise sintática, não semântica.

Além de se propor a resolver problemas de recuperação da informação, como o apresentado acima, a Web Semântica vem colaborando para outras aplicações, tais como a integração de dados, representação e análise do conhecimento, serviços de catalogação, redes sociais, entre outras.

> "A proposta da Web Semântica é estender os princípios dos documentos da web para os dados. Os dados podem ser acessados usando a arquitetura Web (URI por exemplo), e estar relacionados uns com os outros da mesma forma que os documentos já são. Isso significa também criar uma plataforma comum que permita o compartilhamento e a reutilização dos dados por meio das fronteiras das aplicações, empresas e comunidades, podendo ser processados automaticamente por ferramentas, bem como manualmente, inclusive revelando novos relacionamentos possíveis entre porções de dados."Tradução livre do autor de (http://www.w3.org/2001/sw/SW-FAQ)

Segundo a enciclopédia livre, Wikipédia, a palavra semântica, tem origem grega *sēmantiká*, podendo ser traduzida como o estudo do significado, e é utilizando a semântica que Berners-Lee propõe estruturar as informações da Web, com o objetivo de transferir a responsabilidade de interpretação da informação do homem, para a máquina (BREITMAN, 2005). Para isso, os dados da Web precisam estar semanticamente estruturados por meio de uma camada de significado dos dados, os metadados. Os metadados podem ser entendidos como "dados sobre dados", pois trazem informações que descrevem os dados. Na Web Semântica, os metadados são represen-

tados com o modelo de dados RDF, que define um padrão a ser utilizado na codificação XML. Os conceitos de XML e RDF serão explanados nas seções subsequente.

2.3 Metalinguagens

Como já foi citado anteriormente na seção de contextualização histórica, em 1990, Tim Berners-Lee coloca o projeto da *World Wide Web* em prática construindo algumas ferramentas necessárias para seu funcionamento, entre elas uma linguagem de marcação de hipertexto, o HTML (*HyperText Markup Language*).

O HTML trata-se de uma metalinguagem derivada da *Standard Generalized Markup Language* (SGML), uma outra metalinguagem baseada na ideia de que documentos contêm estrutura e outros elementos semânticos que podem ser descritos sem que haja referência à forma como estes elementos são exibidos. Sua sintaxe é baseada em etiquetas (*tags*), palavras encapsuladas pelos sinais '<' e '>'. As *tags* descrevem os dados e comandos necessários para manipulação de um documento na Web e são pré-definidas num conjunto chamado DTD ou *Document Type Definition*, que serve para o navegador como um dicionário das *tags* válidas, seguido do comando a ser executado por cada *tag*. Como cada *browser* já possui o DTD pré-definido, a estrutura do HTML se torna rígida por não aceitar a criação de novas etiquetas. Então, para suprir as limitações do HTML, foi criado o XML (*eXtensible Markup Language*) com o mesmo propósito de descrever o conteúdo semântico e os significados contextuais, além da estrutura e forma de exibição de documentos.

O XML é uma recomendação da W3C[4] (*World Wide Web Consortium*) e se assemelha bastante com o HTML pelo fato de também ser derivado do SGML, logo também utiliza *tags* como padrão de marcação. Mas enquanto o HTML tem como objetivo controlar a forma com que os dados serão apresentados, o XML tem como foco descrever os dados contidos no documento. Desta forma, o XML se torna mais flexível que o HTML devido à possibilidade de adicionar novas *tags* sem a necessidade de alterar o DTD. O interpretador (navegador ou outra aplicação) do XML aceita qualquer tipo de tag, desde que esteja no seu padrão sintático.

O SGML e as linguagens derivadas dele, utilizam uma estrutura baseada em árvore, onde cada *tag* representa um nó, que por sua vez, pode ou não ter outros nós.

Para entender melhor a diferença entre HTML e XML, nos trechos de códigos 2.1 e 2.2 é apresentada, respectivamente, o HTML e o XML.

[4]http://www.w3.org/

Algoritmo 2.1: Exemplo de um código escrito em HTML

```
<body>
    <a href="http://exemplo.com">Exemplo de um link</a>
</body>
```

Algoritmo 2.2: Exemplo de um código escrito em XML para descrever bandas

```
<bandas>
  <banda>
    <id>1</id>
    <nome> Metallica </nome>
    <genero> Rock and roll </genero>
  </banda>
  <banda>
    <id>2</id>
    <nome> U2 </nome>
    <genero> Rock and roll </genero>
  </banda>
</bandas>
```

No codigo 2.1 é mostrado um trecho de código em HTML com duas *tags*: `body` e `a`. A *tag* `body` indica o início (`<body>`) e o fim (`</body>`) do corpo de um documento em HTML. Dentro do corpo do documento, é adicionada a *tag* `a`, criando uma ligação (*link*) para a URL "http://exemplo.com"definida pelo atributo `href`.

Parecido com o HTML, o código 2.2 apresenta um trecho um XML. Na primeira linha é iniciado um nó "bandas", composto por duas bandas (dois nós), cada uma com três nós: id (identificador), nome e gênero. No XML também é possível utilizar atributos nas *tags*, mas, devido à liberdade dada pela linguagem, pode haver interpretação dúbia quanto à utilização de uma propriedade ou de um nó. Desta forma, uma banda poderia ser descrita utilizando somente atributos, como é mostrado no código 2.3:

Algoritmo 2.3: Exemplo de um código escrito em XML mal utilizado

```
<banda id="1" nome="Metallica" genero="Rock and roll"></banda>
```

A abordagem apresentada no código 2.3, apesar de sintaticamente correta, não é considerada uma boa prática, pois as propriedades tem por objetivo descrever o dado (metadado), não representar novos dados. Por isso, a abordagem mais correta para este caso seria a apresentada no código 2.4, onde somente o `id` da banda é uma propriedade, afinal, trata-se de um metadado.

Algoritmo 2.4: Exemplo de um código escrito em XML utilizando boas práticas

```
<banda id="1">
  <nome> Metallica </nome>
  <genero> Rock and roll </genero>
</banda>
```

Segundo Fogg et al. 2001, um *website* lento perde sua credibilidade. A Amazon.com[5] tem um aumento de 1% nas vendas a cada 100 milissegundos que são diminuídos do tempo de carregamento da página (KOHAVI; LONGBOTHAM, 2007). Sendo assim, a velocidade na Internet é um fator decisivo para o sucesso ou fracasso de um negócio. Devida a isso, o XML vem perdendo espaço na Web como uma linguagem para transferência de dados. Sua sintaxe é verbosa, repleta de caracteres de controle que consomem preciosos milissegundos durante uma transferência de arquivo na Internet. Então, um outro formato tem sido utilizado para troca de dados, o JSON (*JavaScript Object Notation*) ou Notação de Objetos *JavaScript*:

JSON é uma formatação leve de troca de dados. Para seres humanos, é fácil de ler e escrever. Para máquinas, é fácil de interpretar e gerar [...] JSON é em formato texto e completamente independente de linguagem, pois usa convenções que são familiares às linguagens C e familiares, incluindo C++, C#, Java, JavaScript, Perl, Python e muitas outras. Estas propriedades fazem com que JSON seja um formato ideal de troca de dados. (CROCKFORD, 2009)

O JSON é constituído de apenas duas estruturas: chave-valor e lista. A estrutura chave-valor, também conhecida como dicionário, define uma estrutura baseada em dados associados à uma chave única e sua sintaxe segue o padrão `{"chave": "valor"}`. A lista ou array, é uma estrutura de dados linear e ordenada, que pode armazenar diferentes tipos de dados, contidos entre cochetes: `["dado1", "dado2"]`. Transcrevendo o código 2.2 em XML para JSON, é possível observar sua característica sucinta, como pode ser visto no codigo 2.5.

Algoritmo 2.5: Exemplo de um código escrito em JSON

```
{ "bandas": [
    {
      "id": 1,
      "nome": "Metallica",
      "genero": "Rock and Roll"
    },
    {
      "id": 2,
      "nome": "U2",
      "genero": "Rock and Roll"
```

[5] http://amazon.com

```
11          } ,
12  ] }
```

Na linha 1 do código 2.5 é iniciado um dicionário com a chave "bandas", que possui como valor uma lista de dicionários. Comparando a quantidade de caracteres dos códigos 2.2 e 2.5, que transmitem a mesma mensagem, a formatação em JSON possui 52 caracteres a menos que os 164 caracteres da formatação em XML, o que representa uma redução de 32%.

2.4 RDF

Devido à interação entre o usuário e a Web provida pela Web 2.0, como foi mostrado na primeira seção deste capítulo, a quantidade de documentos na Web cresceu numa proporção assustadora e desorganizada, dificultando a tarefa de recuperação de informações. Então o RDF surgiu com a proposta de suprir essa deficiência e tornar a Web um ambiente organizado.

De acordo com (Klyne et al. 2004), o RDF (*Resource Description Framework*) é um modelo de dados simples e baseado em grafo direcionado e rotulado nos vértices e arestas, onde os vértices representam os conceitos e as arestas o relacionamento entre conceitos.

O RDF estabelece um padrão de metadados para ser embutido na codificação XML, utilizando a ideia de descrever os dados e os metadados por meio de um esquema de triplas compostas por um sujeito, um predicado e um objeto, simulando a representação de uma frase em linguagem natural.

Gramaticalmente, numa oração, o sujeito é o responsável por realizar ou sofrer uma ação ou estado. O predicado é tudo aquilo que se informa sobre o sujeito e estruturado em torno de um verbo. Por fim, o objeto é o termo da oração que completa o sentido de um verbo. Na frase "Metallica é uma banda de *rock and roll*", Metallica é o sujeito da frase, "é uma banda"o predicado e "de rock and roll"o objeto que complementa o verbo. Na Web Semântica os dados estão estruturados da mesma forma, onde o sujeito é um recurso (identificado por uma URI), o predicado é a propriedade do recurso e o objeto é o valor da propriedade.

Esse modelo permite descrever entidades como pessoas, locais e objetos, de uma maneira simples e flexível, e com um potencial de ligação superior ao de documentos HTML, pois ao invés de interligar apenas documentos, é possível interligar as próprias entidades que, inclusive, podem pertencer a diferentes fontes de dados.

RDF é o modelo de dados adotado como padrão pela W3C em 1999, que tem

> como função representar as informações da Web Semântica, de modo que tais informações possam ser trocadas entre as aplicações, sem que haja perda de sentido (MANOLA & MILLER, 2004 apud SOUZA & AQUINO 2010).

Além da estrutura em triplas, o RDF utiliza padrões de metadados, chamados de *RDF Schema* e definidos por meio de *namespaces*.

Como o RDF não fornece mecanismos para definição de classes e propriedades de um domínio específico, foi desenvolvida uma linguagem para especificação de vocabulário RDF, o *RDF Schema* (BRICKLEY & GUHA, 2004), que funciona como uma extensão semântica do RDF, fornece mecanismos não apenas para definir classes, propriedades dos recursos e tipos dos recursos, como também o relacionamento entre essas propriedades e outros recursos. (BRICKLEY & GUHA, 2004).

Vejamos o código 2.6 de um exemplo de um modelo de dados em RDF para descrever uma banda utilizando triplas e *RDF Schema*.

Algoritmo 2.6: Exemplo de um código escrito em RDF

```
1  <?xml version="1.0" encoding="utf-8" ?>
2  <rdf:RDF
3    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4    xmlns:ns4="http://dbpedia.org/ontology/"
5    xmlns:foaf="http://xmlns.com/foaf/0.1/" >
6    <rdf:type rdf:resource="http://dbpedia.org/ontology/Band" />
7    <rdf:Description rdf:about="http://dbpedia.org/resource/Metallica">
8      <foaf:name xml:lang="en">Metallica</foaf:name>
9      <ns4:genre rdf:resource="http://dbpedia.org/resource/Heavy_metal_music"
           />
10   </rdf:Description>
11 </rdf:RDF>
```

O código 2.6 é baseado em um arquivo RDF retirado da DBpedia[6], por isso os recursos possuem URI com domínio dbpedia.org. O arquivo original foi adaptado neste exemplo para fins didáticos, mas pode ser encontrado na íntegra no apêndice deste trabalho.

A linha 1 especifica a codificação e versão do XML e a linha 2 declara a abertura do bloco `<rdf:RDF>`. Em seguida, as linhas 3 à 5 declaram os atributos xmlns (*XML namespace*), com os prefixos `rdf`, `ns4` e `foaf`. O *namespace* `rdf` possui propriedades padrões e comuns ao modelo RDF, tais como *resource*, *about*, *type*, entre outros. O *namespace* `ns4` é

[6]http://dbpedia.org

um conjunto de propriedades referentes à ontologia em geral, enquanto o `foaf`[7] (*friend of a friend*) define um vocabulário de ontologias para descrever pessoas. A linha 6 indica o tipo do recurso. A linha 7 indica, através da *tag* `<rdf:Description>` qual recurso será descrito, enquanto o *namespace* `rdf:about` indica o recurso, neste caso representado pela URI http://dbpedia.org/ontology/Band. Na linha 8, é utilizado o `foaf` para indicar o nome do recurso e na linha 9 o `ns4` para indicar o gênero musical. Por fim, nas linhas 10 e 11 são definidas as *tags* de encerramento, abertas nas linhas 7 e 2, respectivamente.

Utilizando as informações contidas no exemplo, aplicações que utilizam este modelo podem entender a informação da seguinte forma:

"O documento http://dbpedia.org/resource/Metallica é uma banda cujo o nome é 'Metallica' e pertence ao gênero musical 'http://dbpedia.org/resource/Heavy_metal_music'".

A mesma sentença do exemplo também pode ser representada graficamente utilizando vértices e arestas, respectivamente, recursos e propriedades, como apresentado na figura 5.

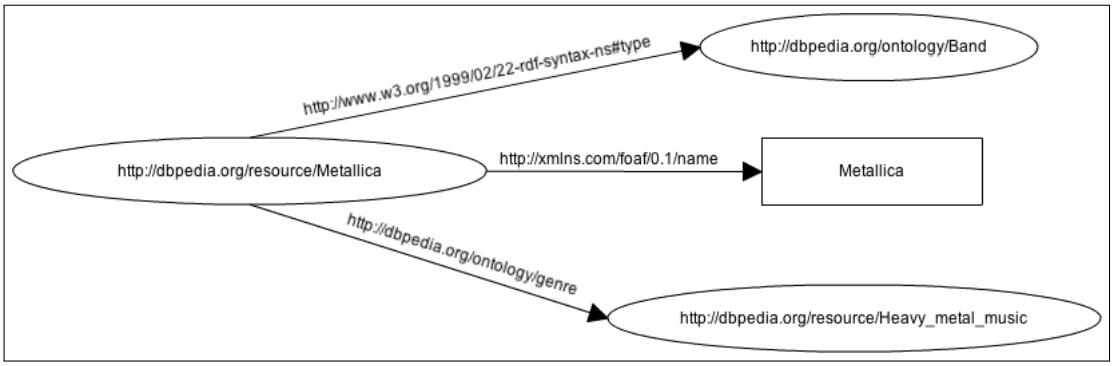

Figura 5: Representação gráfica do modelo em RDF apresentado no código 2.6

2.5 SPARQL e DBpedia

2.6 Linked Data

2.7 Redes complexas

2.8 Ferramentas de Análise

formando uma rede semântica, como pode ser visto no exemplo da figura **??**.

[7] http://xmlns.com/foaf/spec/

3 *Análise e Projeto RDFree*

Com novas soluções tecnológicas para facilitar a geração de conteúdo na internet e torná-la mais rápida e dinâmica, surgiram também novos problemas. Um deles é o problema da organização e recuperação das informações, que estão espalhadas pela internet e à mercê de motores de buscadores para serem recuperadas. Mas a depender da complexidade da condição de recuperação, esses buscadores convencionais têm sua eficácia reduzida, pois seu funcionamento se baseia pela frequência e densidade que os termos buscados aparecem numa determinada página da internet.

Focado no problema da organização e recuperação dos dados, a Web Semântica se propõe a organizar e facilitar a recuperação das informações, adicionando uma nova camada de dados sobre os dados, os metadados. Estes metadados criam uma rede de informação, ligando sujeitos a objetos por meio de predicados, tal como a oração "Metallica é uma banda de Rock", onde "Metallica"é o sujeito, "é uma banda de"é o predicado, e "Rock"é o objeto.

Na Web Semântica, os informações estão interligadas e podem ser representada graficamente como uma rede semântica, onde o sujeito e o objeto são os vértices e o predicado a aresta. Existem no mercado softwares de análise de redes que se propõe a gerar o gráfico de forma personalizada, bem como calcular métricas que caracterizam a rede, como grau de complexidade, clusterização, densidade, e outras. Esses *softwares* funcionam a partir de uma entrada de dados num formato conhecido, que varia de acordo com o *software*.

Apesar de existirem muitas ferramentas de análises de rede, este campo de estudo ainda carece de um modelo computacional que faça algumas etapas preliminares: a de consultar e extrair os dados a partir de uma base de dados e depois torná-los legíveis para estas ferramentas.

Portanto, este trabalho tem como objetivo desenvolver uma ferramenta de extração e conversão de dados RDF para um formato interpretável por *softwares* de análise de redes. Esta ferramenta foi denominada RDFree, nome proveniente da junção entre o acrônimo RDF e a palavra em inglês *free* (livre).

Com a utilização do RDFree, pesquisadores da área de redes semânticas podem gerar fa-

cilmente o gráfico e as métricas da rede, a partir dos dados presentes na *Linked Data*, e utilizar esses dados como subsídio para suas pesquisas.

Nas próximas seções será apresentada a ferramenta de conversão desenvolvida, dando uma visão mais aprofundada sobre o seu funcionamento e as etapas que foram necessárias para o alcance do objetivo fulcral desde trabalho. [sumarizar os itens das próximas seções]

Para nortear este projeto, foram analisados trabalhos relacionados que o oferecessem recursos de consulta SPARQL ou geração de gráficos de rede. Na próxima seção, cada projeto é apresentado e comparado com os demais em uma tabela.

3.1 Trabalhos Relacionados

Na busca por soluções semelhantes ao objetivo deste trabalho, que pudessem contribuir na especificação do RDFree, foram encontrados alguns projetos no domínio da web semântica e redes complexas. Esta pesquisa contribuiu para concepção do RDFree e confirmou a deficiência por ferramentas de extração e conversão de dados para um formato reconhecível por softwares de análise de redes.

Uma das mais utilizadas ferramentas de análise de redes é o *software open-source* Gephi, que foi apelidado pela comunidade como o *Photoshop*[8] para gráficos. Segundo a descrição feita pelo site oficial[9], o Gephi é uma ferramenta de visualização interativa e uma plataforma de exploração de todos os tipos de redes e sistemas complexos, dinâmicos e gráficos hierárquicos, capaz de ser executado nos sistemas operacionais Windows, Linux e Mac OS X.

O Gephi, além de suas ferramentas nativas, possui ainda alguns *plugin*[10], como é o caso do SemanticWebImport, um plugin desenvolvido pela Wimmics[11], que permite a importação de dados semânticos para o Gephi. Os dados importados são obtidos a partir de uma consulta SPARQL em uma base dados semântica (RDF).

Diferente do RDFree, que tem o propósito de ser uma ferramenta genérica e modularizada, o SemanticWebImport é limitado ao Gephi, sendo incapaz de funcionar de forma independente ou com outras ferramentas. Outro ponto a ser observado na pesquisa é que o resultado obtido com este plugin não foi o esperado. A partir de uma mesma consulta SPARQL a base de dados da DBpedia, os resultados obtidos com o SemanticWebImport não foram compatíveis

[8] Software líder do mercado dos editores de imagem
[9] https://gephi.org
[10] Complementos que podem ser incorporados à ferramenta
[11] http://wimmics.inria.fr/

com os resultados obtidos a partir de uma consulta direta à base de dados por meio do Virtuoso SPARQL[12]. Essa inconsistência dos dados não é encontrada no RDFree, que obteve resultados idênticos ao esperado.

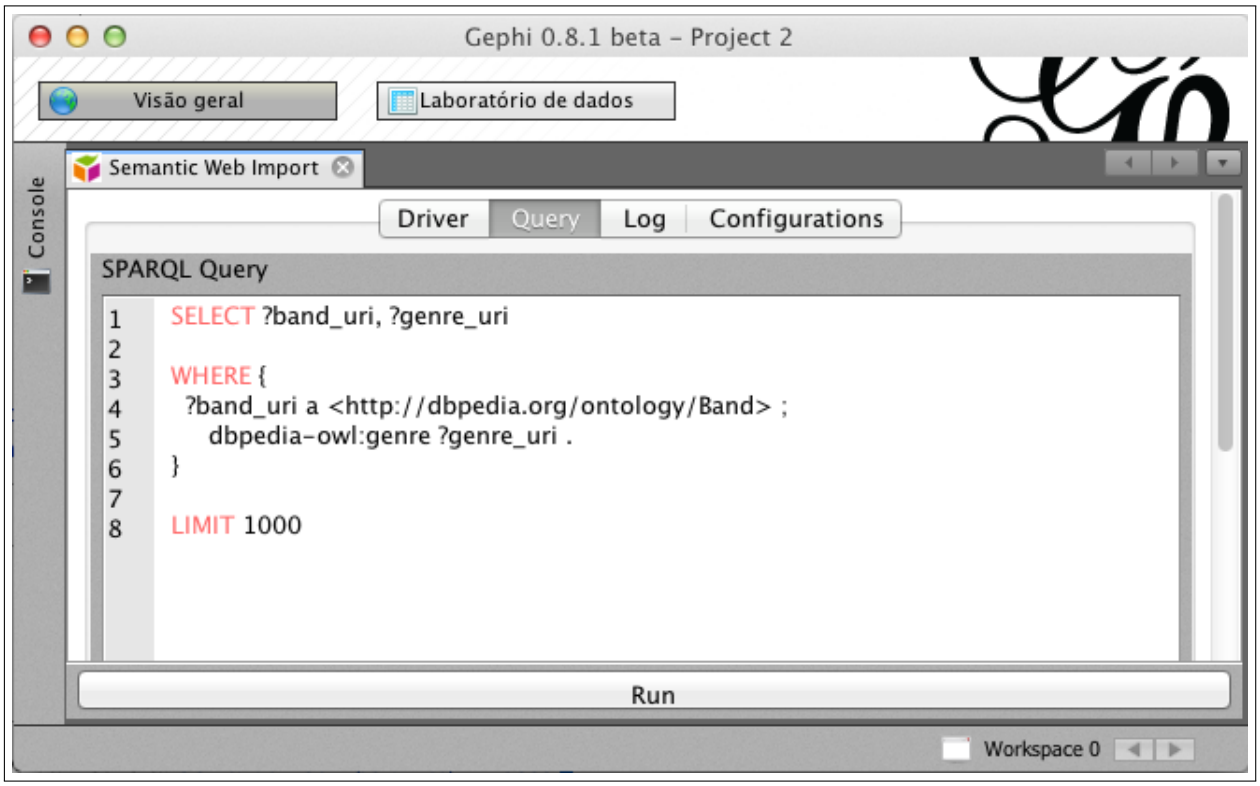

Figura 6: Janela do SemanticWebImport com consulta SPARQL

Uma ferramenta muito completa encontrada neste estudo foi a *TopBraid Composer*. Implementada como um plugin do Eclipse, ela serve como um ambiente de modelagem e edição de ontologias, construção de aplicações semântica, criação de gráficos e ainda, como o site oficial o descreve, "a melhor ferramenta SPARQL no mercado". O *TopBraid Composer* é distribuído em três versões: Free, Standard e Maestro, sendo esta última a mais completa e licenciada por U$3.450,00. Contudo, este software está mais voltado para criação e modelagem de ontologias, sendo incapaz de fazer uma consulta remota e converter os dados para um formato desejado.

Mais simples que a *TopBraid Composer*, o software Welkin, open-source desenvolvido em 2006 no MIT (*Massachusetts Institute of Technology*), tem o objetivo de ser um visualizador de modelos RDF. Seu *input* está limitado a arquivos do tipo RDF, RDFS, OWL, n3 ou turtle, não sendo capaz de realizar consultas SPARQL em base de dados remotas ou locais. Suas funcionalidades também são reduzidas, limitando-se à criação do gráfico da rede e dos gráficos de grau de entrada, grau de saída e coeficiente de clusterização, como é possível ver na figura 7.

[12]Interface utilizada pela DBpedia para fazer consultas SPARQL

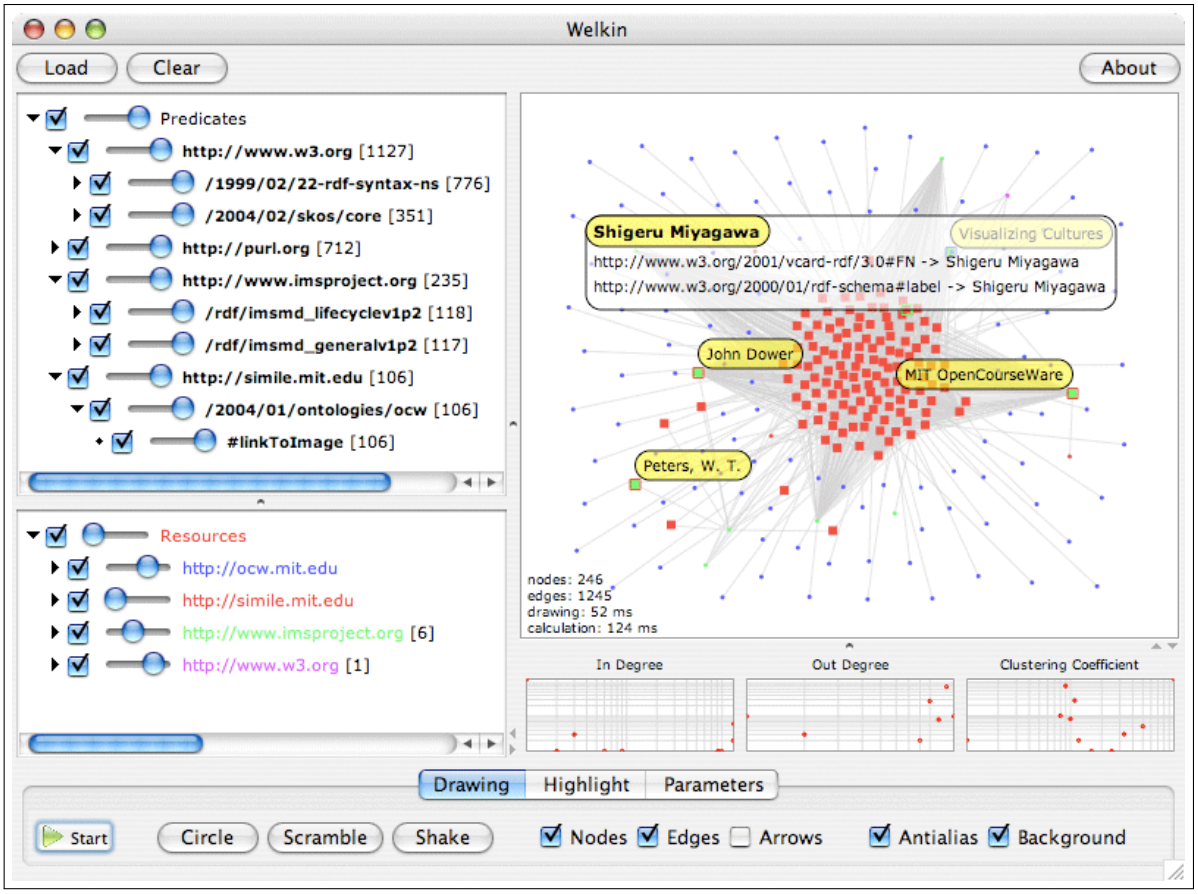

Figura 7: Gráfico gerado pelo Welkin

Assim como o *Welkin*, outra ferramentas se propõe a ser um visualizador de modelo RDF, tais como *visualRDF*, *Visual Browser*, *RDF Gravity* e *IsaViz*. Todas estas encontram-se desatualizadas e com última versão datada de 2004 à 2007.

Vale citar ainda a *Force-Directed Graph*, uma ferramenta online escrita em *JavaScript* e que, entre todas as ferramentas estudadas, é a que se encontra mais atualizada (11/2012). Seu objetivo limita-se à representação gráfica da rede a partir de uma entrada de dados padrão no formato JSON. Apesar do seu propósito estar mais próximo do *Gephi* e mais distante do objetivo deste trabalho, o *Force-Directed Graph* serviu como inspiração para modelagem do RDFree, que também utiliza um arquivo JSON como normalizador dos dados.

Concluída a pesquisa por trabalhos correlatos, foi feita uma comparação entre eles e o RDFree, como observado na tabela 1. Em seguida, na próxima seção são descritos os requisitos que nortearam a concepção do RDFree.

Tabela 1: Tabela comparativa entre trabalhos relacionados

	Open Source	Consulta SPARQL	Gera gráfico	Exporta para outros formatos
Semantic Web Import		x		
TopBraid Composer		x	x	
Welkin	x		x	
Force-Directed Graph	x		x	
RDFree	x	x		x

3.2 Requisitos e Projeto Arquitetural

A análise de trabalhos correlatos foi importante no processo de levantamento de requisitos funcionais (RF) e não funcionais (RNF) do RDFree, como pode ser visto na tabela 2.

Tabela 2: Requisitos funcionais e não funcionais

Requisito	Descrição
RF1	A aplicação deve conseguir importar dados de uma consulta SPARQL a uma base de dados RDF remota ou local.
RF2	Os dados devem ser preparados e normalizados para um formato padrão.
RF3	Os dados normalizados e convertidos devem ser salvos em um arquivo.
RNF1	As definições de configuração do conversor devem estar em um arquivo separado do código.
RNF2	A aplicação deve dar suporte a inclusão de novos módulos de tradução.

O projeto arquitetural deste trabalho é ilustrado pela figura 8 em forma de fluxograma, mostrando as etapas realizadas pela aplicação desde a sua inicialização, até a criação do arquivo de saída.

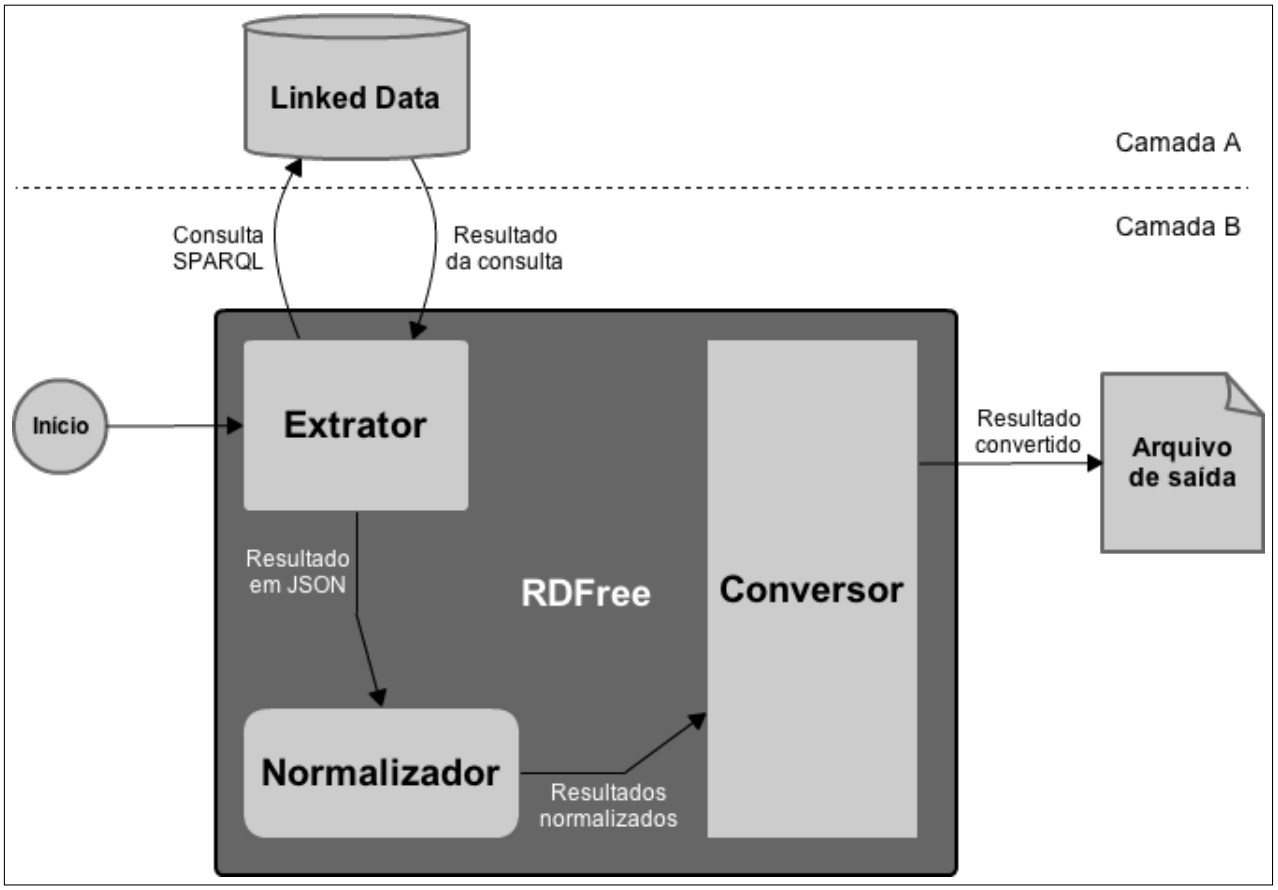

Figura 8: Arquitetura do RDFree

Na camada (A) está a base de dados RDF, que pode ser remota ou local e que tem a função de retornar o resultado de uma consulta para a camada (B), onde estão os três módulos que compõe a ferramenta RDFree:

- **Extrator** - é o módulo responsável por se comunicar com a base de dados, executar a consulta e importar os resultados para a aplicação, contemplando o requisito RF1.
- **Normalizador** - executa passos comuns à todos os conversores, criando uma estrutura de dados já com os nós e arestas da rede, satisfazendo o requisito RF2.
- **Conversor** - este módulo funciona como um tradutor, recebendo como entrada os dados padronizados, gerados pelo normalizador, e retornando um formato de acordo com sua espeficicação. Atende aos requisitos RF3 e RNF2.

Para compreender melhor o funcionamento e interação desses módulos, na próxima seção é mostrada a solução computacional utilizada na criação de cada módulo, apresentando seu funcionamento de forma mais detalhada, bem como as tecnologias empregadas no projeto.

3.3 Codificação da Solução

O RDFree foi desenvolvido utilizando a linguagem de programação Python, portanto, os trechos de códigos encontrados nesta seção, bem como as bibliotecas utilizadas, estão escritos nesta mesma linguagem.

O módulo de extração inicia a execução do RDFree e é responsável por importar os dados que serão convertidos. A conexão com o endpoint e execução da consulta SPARQL são feitas com o auxílio do framework SPARQLWrapper, que provê uma interface amigável e simples para o programador, como pode ser visto no código 3.1.

Algoritmo 3.1: Exemplo de SPARQLWrapper

```
1 sparql = SPARQLWrapper("http://dbpedia.org/sparql")
2 sparql.setQuery("SELECT * WHERE { ?s ?p ?o } LIMIT 10")
3 sparql.setReturnFormat(JSON)
4 results = sparql.query().convert()
```

No código 3.1, na primeira linha é instanciado um objeto SPARQLWrapper que fará uma consulta à DBpedia. Na linha 2 é determinada a consulta SPARQL que será realizada. Esta consulta retorna o sujeito, predicado e objeto dos dez primeiros resultados. Na linha 3 é determinado o formato dos dados recebidos, neste caso é utilizado o JSON por ser de fácil manipulação no Python. Por fim, na linha 4, a consulta é realizada e os resultados convertidos para JSON.

Ao receber os resultados da consulta, o módulo de extração converte os resultados para o formato JSON, que é um formato de fácil manipulação no Python.

Após a extração, o resultado é normalizado para um formato genérico, que auxiliará na próxima etapa de conversão. Este novo formato, também em JSON, é composto por duas listas: uma de vértices ou nós e outra de arestas. A lista dos vértices é formada por todos os recursos (sujeitos e objetos) únicos presentes na consulta e seguidos de um identificador numérico, que será usado na lista de arestas, formada por pares de relacionamento entre os vértices.

No exemplo de uma consulta pelos filhos de Francisco, o resultado pode ser o mostrado na tabela 3 e o JSON desse resultado representado no código 3.2.

Tabela 3: Exemplo de resultado de uma consulta

Pai	Filho
Francisco	Mirosmar
Francisco	Emival

Algoritmo 3.2: Estrutura normalizada

```
# Vertices
{
  "pai" : {
    "Francisco" : {"id" : 1}
  },
  "filho" : {
    "Mirosmar" : {"id" : 2},
    "Emival" : {"id" : 3},
  }
}

# Arestas
[
  [1, 2],
  [1, 3]
]
```

O módulo de conversão é responsável por transformar os dados normalizados para um padrão interpretável por ferramentas de análise de rede. Até a conclusão deste trabalho, foi criado apenas o módulo de conversão para o formato GEXF (Graph Exchange XML Format), uma linguagem criada em 2007 juntamente com o Gephi, com o objetivo de padronizar a descrição de estruturas complexas de redes, e portanto, o formato de entrada utilizado pelo Gephi. O projeto GEXF possui ainda bibliotecas para Python, Java, C++, JavaScript e Perl que facilitam na criação do arquivo.

3.4 Aplicando a Solução em um Domínio

Para mostrar o funcionamento da ferramenta desenvolvida neste trabalho, será utilizada a base de dados da DBpedia em Português, para buscar por bandas de música e seu respectivo gênero musical. Para fins didáticos, a consulta será limitada a mil resultados, que é equivalente a mil relacionamentos. A consulta SPARQL utilizada para definir o domínio estudado pode ser

visto no código 3.3.

Algoritmo 3.3: SPARQL utilizado para consultar bandas e seus gêneros

```
SELECT ?band_uri, ?genre_uri
WHERE {
    ?band_uri a <http://dbpedia.org/ontology/Band> ;
    dbpedia-owl:genre ?genre_uri .
}
LIMIT 1000
```

Neste exemplo, será utilizado o Gephi para gerar a representação gráfica dos dados, por isso o arquivo gerado será no formato GEXF.

A consulta SPARQL mostrada no codigo 3.3, bem como a URI da DBpedia devem estar no arquivo de configuração, como o mostrado no codigo 3.4. Feito isso, a aplicação já pode ser iniciada executando o arquivo main.py, que salvará o resultado no arquivo output.gexf.

Algoritmo 3.4: Arquivo de configuração

```
OUTPUT_FILE = 'output.gexf'
ENDPOINT = "http://dbpedia.org/sparql"
QUERY = '''
  SELECT ?band_uri, ?genre_uri
  WHERE {
    ?band_uri a <http://dbpedia.org/ontology/Band> ;
    dbpedia-owl:genre ?genre_uri .
  }
  LIMIT 1000
'''

# GEXF
CREATOR = 'Shankar Cabus de Teive e Argollo'
DESCRIPTION = u'Aplicando a solucao em um dominio'
DEFAULTEDGETYPE = 'undirected'
MODE = 'static'
LABEL = 'Relacionamento entre bandas e genero'
```

Abrindo o arquivo output.gexf no Gephi é possível encontrar o grau médio da rede (1,727), densidade do grafo (0.003), modularidade (0,667), entre outras métricas. Ao abrir o Gephi, já é exibido um gráfico gerado de forma aleatória, mas para efeito didático, na figura 9 é mostrado o gráfico da rede distribuído utilizando o algoritmo de Fruchterman Reingold e classificado por excentricidade, onde os pontos vermelhos são os gêneros musicais e os azuis são as bandas.

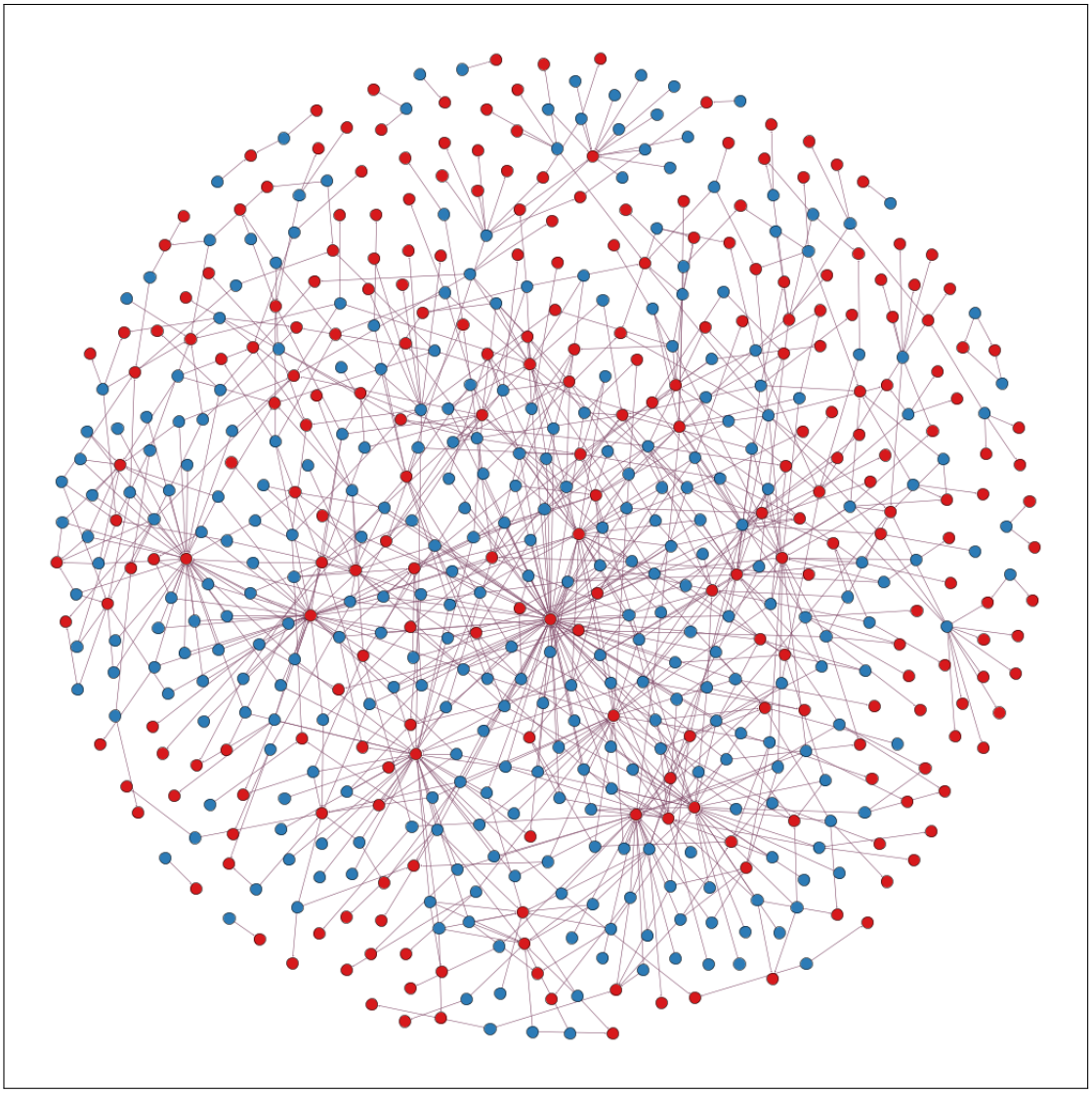

Figura 9: Representação gráfica da rede semântica gerada pelo Gephi

3.5 Análise da solução

Conforme foi mostrado na seção 3.4, o RDFree atendeu à todos os requisitos definidos neste trabalho, executando a consulta a uma base de dados RDF e exportando os dados para um formato aceito por uma ferramenta de análise de redes.

Contudo, o RDFree carece de uma interface amigável para usuários que não tenham conhecimento sobre a linguagem de programação utilizada no desenvolvimento do projeto.

Referências

. . . Disponível em: <.>. Acesso em: .

BERNERS-LEE, T. *Tim Berners-Lee*. 1991. Disponível em: <https://groups.google.com/group/alt.hypertext/tree/browse_frm/thread/7824e490ea164c06/f61c1ef93d2a8 Acesso em: 30 de novembro 2012.

BLATTMANN, U.; SILVA, F. C. C. da. Colaborac?a?o e interac?a?o na web 2.0 e biblioteca 2.0. 2007.

BORCHANI, V. Los centros de recursos para el aprendizaje y las nuevas tecnologi?as de informacio?n e comunicacio?n en la biblioteca universitaria. in. 2007.

BREITMAN, K. Web sema?ntica: A internet do futuro. In: ____. *Web Sema?ntica: A Internet do Futuro*. [S.l.: s.n.], 2005.

CONNOLLY, D. *A Little History of the World Wide Web*. 2000. Disponível em: <http://www.w3.org/History.html>. Acesso em: 30 de novembro 2012.

CROCKFORD, D. *Introdução ao JSON*. 2009. Disponível em: <http://json.org/json-pt.html>. Acesso em: 4 de dezembro de 2012.

KOHAVI, R.; LONGBOTHAM, R. Online experiments: Lessons learned. *The Amazon statistic was taken from a presentation by Greg Linden at Stanford*, 2007.