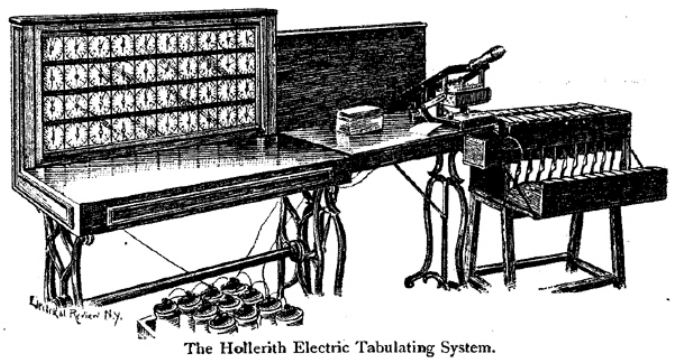
1
Introdução

Nos últimos anos, a Web Semântica vem se destacando como padrão para publicação e troca de informações na internet, uma vez que, oferece um framework comum que permite o compartilhamento e reutilização de dados entre usuários. Além disso, oferece tecnologias que possibilitam descrever, modelar e consultar informações. Com a adoção destes padrões, provedores de conteúdo podem publicar informações utilizando-se vocabulários específicos de domínio e *interfaces* de consulta, facilitando o acesso aos dados.

A World Wide Web Consortium (W3C) recomenda a utilização do padrão Linked Data [1] na representação de dados abertos. Este padrão baseia-se na representação de dados na forma de um conjunto de triplas RDF. Neste cenário a conversão de conjuntos de informações (esquemas de banco de dados e suas instâncias) para conjuntos de dados RDF é um passo importante.

A quantidade de dados convertidos para RDF cresce significativamente a cada dia [2][3][4]. No entanto, embora seja possível observar um crescimento exponencial no tamanho da Web como um todo, ainda há uma diferença significativa entre o crescimento da Web e o crescimento da Web Semântica [4]. Acredita-se que esta diferença seja devida à falta de ferramentas para gerar dados RDF, a partir da conversão de dados em diferentes formatos.

FIG. 2.



The Hollerith Electric Tabulating System.

Figura 1. Sistema eletrônico de tabulação desenvolvido por Hollerith em 1989, a fim de facilitar o censo Norte Americano.

É interessante observar que a Web Semântica vem crescendo e sendo impulsionada pelo mesmo setor que apoiou o surgimento dos primeiros computadores, o setor governamental. A máquina de Hollerith, ilustrada na Figura 1, foi utilizada para auxiliar o censo Norte Americano em 1890, reduzindo o tempo de processamento de dados de sete anos, do censo anterior, para apenas dois anos e meio e ajudando o governo a economizar cinco milhões de dólares [5]. Curiosamente, os mesmos objetivos que levaram Hollerith à automatização do censo Americano, exemplificados pelo trecho abaixo, estão levando os governos à adoção de tal padrão. Novamente, cientistas estão somando forças no intuito de desenvolver mecanismos que melhorem o acesso e processamento da informação disponível.

"se o número de desempregados meses forem devidamente enumerados e compilados com referência à idade, ocupação, etc, muita informação de grande valor pode ser obtida para o estudioso de problemas econômicos que afetam nossos assalariados"[6] Hermann Hollerich, 1889

Neste trabalho, propomos Babel, um framework de apoio à conversão de dados armazenados em formatos tradicionais, e.g. bancos de dados relacionais e planilhas, para formatos semanticamente ricos, e.g., RDF-XML, RDF-Ntriples e RDFa. Mais do que um simples conversor, Babel facilita a criação de mapeamentos e a reutilização de vocabulários através de templates. Além disso, fornece uma API única, que possibilita mesclar e publicar informações de várias fontes de dados em diferentes formatos, o que é fundamental para assegurar a adoção do padrão Linked Data.

1.1 Motivação

É incrível como ainda estejamos tentando resolver o mesmo problema de Hollerith no final do século 19. É bem verdade que o objeto mudou. Hollerith em 1889 estudava um mecanismo que possibilitasse a computabilidade do Censo Norte Americano, que em síntese são informações coletadas de indivíduos. A Web Semântica, por outro lado, estuda mecanismos que possibilitem a

interligação de grandes conjuntos de dados, oriundos de bases diferentes, mas dada as devidas proporções, ainda são nossas informações.

Muitas conquistas e avanços foram alcançados nos mais de 100 anos que se passaram e a interligação de dados ainda parece ser um problema difícil de resolver. Esta dificuldade deve-se ao fato de que embora a tecnologia possibilite um processamento cada vez maior de informações, ela também possibilita que um número cada vez maior de dados possam ser produzidos, o que pode ser facilmente constatado pelo constante crescimento de dados publicados na Web.

Nos últimos anos verificou-se um aumento de interesse em ferramentas de apoio ao processo de publicação de dados na Web, principalmente na esfera governamental[12]. Segundo Tim Berners-Lee, a busca na publicação de informações governamentais é motivada basicamente por três razões [7]:

- *Aumentar a consciência cidadã nas funções do governo para permitir o engajamento;*
- *Contribuir com informações valiosas sobre o mundo;*
- *Permitir que o governo, o país e o mundo funcionem com maior eficiência.*

No entanto, a dificuldade para utilização dos dados governamentais demonstrou que publicá-los na Web era muito mais que colocar informações em sites, para que pudessem ser acessadas publicamente. Embora o padrão Linked Open Data tenha demonstrado ser um caminho importante para estas instituições, como veremos, ainda existem barreiras que impedem sua difusão.

Para Auer [4], a razão principal é a existência de ferramentas que se perdem na complexidade da geração de mapeamentos para banco de dados, por três razões:

- *Identificação de dados privados e públicos;*
- *Uso apropriado dos vocabulários existentes;*
- *Perda da descrição original do esquema do banco de dados.*

[1] www.data.gov.uk – site de publicação de conjunto de dados do governo britânico.

[2] www.data.gov – site de publicação de conjunto de dados do governo norte americano.

Butler [8] sintetiza os obstáculos para a adoção dos padrões da Web Semântica como sendo:

- *A complexidade do RDF/XML;*
- *Uso indevido de ferramentas complexas;*
- *Suporte a múltiplas versões de vocabulários;*
- *Suporte a múltiplos vocabulários;*
- *Simplificar a criação, validação e processamento da meta-informação da Web Semântica;*
- *Ocultar a complexidade da Web Semântica de usuários leigos;*
- *Prover técnicas de mapeamento entre múltiplos vocabulários e versões;*
- *Padronização de Vocabulários;*
- *Problemas de imprecisão ou inconsistência.*

Há várias razões para que indivíduos, empresas, instituições públicas ou privadas estejam interessadas em publicar suas informações de uma forma acessível e interligada. Pessoas desejam divulgar suas informações e atividades; Empresas desejam que usuários tenham acesso aos seus produtos e serviços; Instituições públicas querem tornar a sua administração transparente; Instituições privadas, interligar suas bases de conhecimento. Estas são algumas razões pelas quais a Web Semântica está tomando um lugar de destaque no cenário atual, isso porque, dentre outras coisas, ela possibilita que informações, antes inacessíveis, possam ser processadas, e vinculadas a um conjunto de outras informações armazenadas em repositórios distintos.

Em paralelo ao contínuo desenvolvimento da Web de hipertexto, testemunhamos um rápido crescimento da Web Dados, onde muitas empresas (Google, Flicker, Facebook, Amazon e outras) começaram a doponibilizar o seu conteúdo através de APIs, com o objetivo de disseminar conteúdo em RDF de uma maneira interligada[3]. Também assistimos ao

[3] http://www.linkeddata.org

lançamento do RDFa, tecnologia que permite o acesso e consumo de páginas HTML como fontes de dados estruturados. Apesar de disponível, esta riqueza de informações não é acessível através das formas de busca tradicionais porque estão disponíveis de uma forma não padronizada, em estruturas que variam em forma e conteúdo. É exatamente neste ponto que a técnica de Mashup pode ajudar, agrupando dados de diferentes bases, através do alinhamento de seus vocabulários, permittindo a compreensão do conjunto dos dados. Embora alguns editores de mashups tenham sido propostos (como o Google Mashups, Popfly da Microsoft, IBM sMash, e Yahoo Pipes), sua criação ainda depende de programadores habilidosos.

O estado atual da Linked Open Data sugere que interligar a informação não é um processo trivial (apenas 11,5% da informação disponível na LOD está interligada, Tabela 1). Mais ainda, estudos indicam que a Surface Web cresce em escala de bilhares a cada ano, sem contar a gigantesca quantidade de informação que se encontra ainda inacessível na Deep Web.

"At the time of writing there are fifty-three million blogs on the Internet, and this number is doubling every six months."

"YouTube, is a portal of amateur videos that, at the time of writing, was the world's fastest-growing site, attracting sixty-five thousand new videos daily…"

Andrew Keen [9]

Esses números demonstram o imenso abismo que ainda nos separa do modelo da Web de Dados proposto por Tim Beners Lee. Não porque a informação não possa se tornar de fato legível para as máquinas, mas por que ainda não dispomos de mecanismos eficazes que permitam processá-la e interligá-la. Porém, sem sombra de dúvida, a Web Semântica está facilitando o acesso à informação de uma maneira que ela pode ser interligada por terceiros para ser analisada e processada das mais variadas formas.

As barreiras para difusão da Web Semântica se tornam ainda mais evidentes quando se comparamos a diferença existente entre o crescimento da Web

propriamente dita e a Web Semântica, que pode ser constatada nos índices provenientes do Swoogle [4] [10] e do World Wide Web Size [11].

A World Wide Web Size é um site criado para estimar o tamanho da WWW baseado no número de páginas indexadas pelas máquinas de busca Google, Bing, Yahoo Search e Ask. O índice é uma estimativa, calculada a partir da subtração da sobra dos índices das quatro máquinas de busca obtidos em seqüência. São apresentados dois índices, dentre os possíveis: um partindo do Yahoo – YGBA – e o outro partindo do Google – GYBA.

O Gráfico 1 demonstra que, embora o crescimento da Web Semântica tenha sido grande nos últimos anos, o número de documentos indexados pela ferramenta de busca Swoogle ainda é tímido. A comparação do número de documentos encontrados pelo Swoogle com o número de documentos encontrado pelo World Wide Web Size, utilizando o índice GYBA durante os dois últimos anos, sugere que o crescimento da Web Semântica vem assumindo um comportamento linear em comparação ao crescimento da Web, que vem tendo um comportamento exponencial.

Em Janeiro de 2011, o número de documentos na Web, estimado pelo site World Wide Web Size, utilizando-se o menor índice encontrado em Janeiro de 2011, foi de aproximadamente 18.86 bilhões. Em contraste, o Swoogle indexou 1.32 milhões de documentos e 1.11 bilhões de triplas no mesmo período, uma média de 840 triplas por documento indexado. Analisando o estado atual dos *datasets* da Linked Open Data é possível encontrar um total de 35 bilhões de triplas [3]. Por analogia, podemos estimar que estas triplas estejam distribuídas em aproximadamente 42 milhões de documentos; mais ainda, supor que a maioria dos documentos publicados no formato RDF está concentrada em grandes *datasets*, a exemplo dos conjuntos de dados governamentais que detém 40% de toda a informação publicada na LOD [2].

[4]Swoogle - um mecanismo de busca para documentos semânticos – ontologias – na Web.

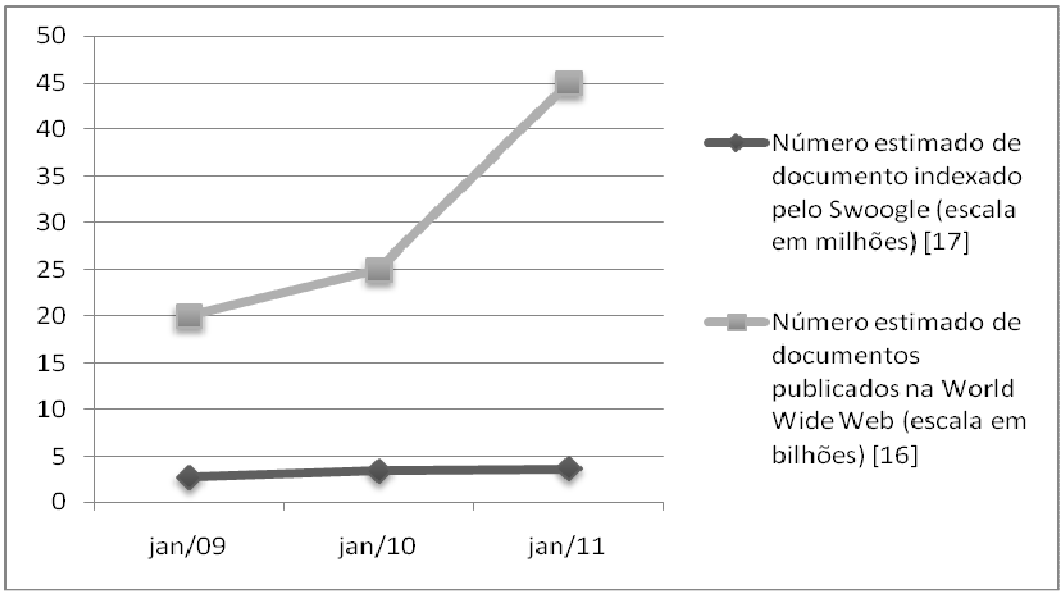


Gráfico 1. Crescimento estimado Web e da Web Semântica nos últimos dois anos, utilizando o índice GYBA do World Wide Web Size e o índice de documentos indexados pelo Swoogle.

Também é possível verificar, na Tabela 1, que uma pequena parcela da informação publicada na LOD está interligada, cerca de 11,5%. De todos os conjuntos de dados publicados, destaca-se o conjunto de Ciências da Vida, que possui um montante de 38% de dados interligados, mas, no entanto, grande parte deste conjunto está restrita a dados genéticos como grandes cadeias de DNA. Dessa forma, ainda não é possível afirmar qual será o futuro da Linked Data, os vários movimentos e técnicas de publicação de dados ainda não conseguiram um volume considerável de interligação, o que nos faz questionar a sua possibilidade real. De qualquer forma, mesmo que não seja possível a publicação total dos dados de uma forma interligada, os formatos da Web Semântica permitem que eles possam ser utilizados e interligados por terceiros.

Domínios	Número de Conjunto de Dados	Triplas	%	Links Externos	%	% Links Externos na LOD
Mídia	25	1.841.852.061	5,82	50.440.705	10,01	0,582582
Geografia	31	6.145.532.484	19,43	35.812.328	7,11	1,381473
Governo	49	13.150.009.400	42,09	19.343.519	3,84	1,616256
Publicações	87	2.950.720.693	9,33	139.925.218	27,76	2,590008
Conteúdo entre domínios	41	4.184.635.715	13,23	63.183.065	12,54	1,659042
Biologia	41	3.036.336.004	9,6	191.844.090	38,06	3,65376
Conteúdo gerado por usuário	20	134.127.413	0,42	3.449.143	0,68	0,002856
Total	295	31.634.213.770		503.998.829		11,485977

Tabela 1. Número de conjunto de dados distribuídos por domínio na LOD [2] em Setembro de 2011.

Outro dado nos mostra que a LOD não é tão grande como imaginamos. A DBPedia[5] contém 1 bilhão de triplas distribuídas em 22 *gigabytes*[6] no formato N-Triples e N-Quad. Sabendo que há duplicidade de informação, ou seja, cada declaração RDF no formato N-Triples está reescrita no formato N-Quad, podemos estimar que o tamanho real da DBpedia seja de aproximadamente 11 GB, o que daria à LOD um tamanho total de aproximadamente 348 gigabytes. Segundo esta constatação, a LOD, poderia ser armazenada em qualquer computador pessoal, o que nos mostra certa fragilidade.

Além disso, Auer [4] constatou que a maioria dos termos indexados pelas ferramentas de busca como Swoogle, estão em RDF, RDFS, OWL e se utilizam de vocabulários RDF populares tais como FOAF, DC, RSS.

[5] www.dbpedia.org – Trata-se de um dos maiores e principais *datasets* da LOD que é um esforço comunitário para extrair os dados da Wikipedia e torná-los acessíveis.

[6] http://downloads.dbpedia.org/3.7/ - D*ataset* da DBpedia 3.7 disponível para download.

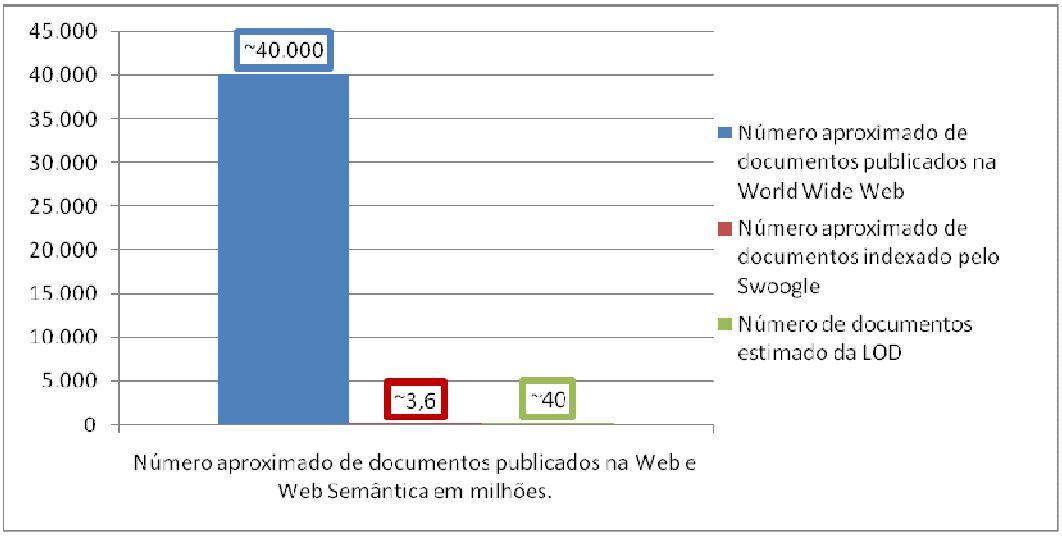


Gráfico 2. Comparação entre o tamanho da World Wide Web e a Web Semântica em 2011 utilizando os índices aproximados do World Wide Web Size e Swoogle encontrados em Janeiro de 2011, com o tamanho estimado de documentos na LOD (Uma estimativa utilizando a relação do número de triplas por documento extraído do Swoogle e aplicado à LOD).

Outro aspecto que merece nossa atenção é o desempenho das ferramentas de publicação de Linked Open Data, muito inferior a padrões e aplicações já estabelecidos como banco de dados relacionais. Há muitos trabalhos que recomendam a utilização de banco de dados relacionais para armazenar e consultar informações em RDF [12-15]. Embora haja muitos estudos envolvidos na transformação da álgebra SPARQL em SQL [12-13], ainda não há uma maneira definitiva e estabelecida para a realização desse processo. Dentre os fatos que impedem uma transformação eficiente da álgebra SPARQL para SQL podemos destacar a diferença de expressividade contida nas linguagens, a álgebra SPARQL é muito mais rica que a álgebra SQL [16]. Os resultados de um teste, para avaliar o desempenho das aplicações, ilustradas pelo Gráfico 3, sugeriu que conversores simples possuem um desempenho muito superior ao apresentado pelas ferramentas que manipulam o grafo RDF, o que leva a acreditar que, uma vez que as questões envolvendo a conversão de SPARQL para SQL sejam resolvidas, o desempenho destas aplicações deve melhorar. A seguir reproduzimos comentários de pesquisadores da área que servem para ilustrar este fato.

“Unfortunately, the join rule stated above does not fully reproduce SPARQL semantics"

Richard Cyganiak [12]

“a more sophisticated algorithm is required to express nested optional graph patterns"

Stephen Harris [13]

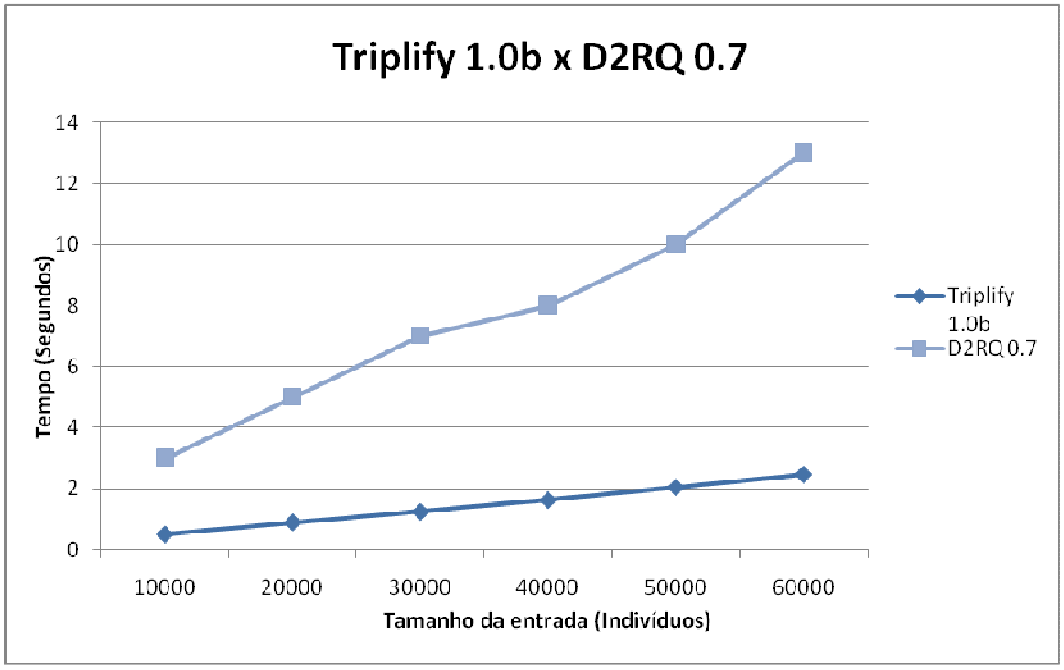


Gráfico 3. Comparação de desempenho entre Triplify 1.0b e D2RQ 0.7 na geração de indivíduos do tipo FOAF contendo apenas um atributo nome.

Por outro lado, as informações provenientes do Euro Semantic Web [17], demonstram que poucas destas ferramentas são, de fato, extensíveis. Os 73 conversores catalogados dão suporte a 49 tipos de formatos distintos. Em resumo, há uma ferramenta e meia para cada formato, o que é um número relativamente baixo. A ferramenta que é mais extensível apresentou suporte a cinco formatos, e a menos extensível apenas a um. Bergman [18] estima que a combinação de dados disponíveis na Deep Web e Surface Web seja superior a meio trilhão de documentos. Os números indicam a necessidade premente de ferramentas, métodos e técnicas para aprimorar e melhorar o acesso a tais informações, um processo também conhecido como *surface* [19].

Acredita-se que a falta de ferramentas que facilitam a conversão de dados armazenados em bancos de dados relacionais, planilhas e outras fontes, é um fator

de impacto na publicação de conteúdo semanticamente enriquecido na Web. Apesar do vasto número de soluções que suportam a conversão de diferentes tipos de dados em RDF – Triplify [4], D2RQ [20] e Virtuoso [21-22] – ainda há problemas que inibem o desenvolvimento da Web Semântica. Algumas soluções [20] mostram uma mudança de atitude por parte dos pesquisadores, tornando o processo de conversão mais simples e acessível para os usuários, embora, em sua grande maioria, possam ser notadamente constatados problemas de extensibilidade, escalabilidade ou mesmo usabilidade. As principais ferramentas de conversão de dados RDB em RDF [4,20-22] não utilizam a mesma linguagem de mapeamento. É preciso simplificar e profissionalizar o processo. Algumas iniciativas vêm sendo tomadas para o estabelecimento de padrões, a exemplo do RDB2RDF [23], movimento que visa estabelecer um padrão para o mapeamento de dados no formato relacional para o RDF, no entanto, ainda há muito que ser feito.

A adesão aos *Wrappers* vem sendo grande, por serem simples e de fácil uso. Nos *Wrappers* toda a conversão é realizada utilizando mapeamento direto (*Direct Mapping*), onde as propriedades, classes e valores são extraídos e criados a partir da fonte original. É importante notar que o mapeamento das propriedades, valores e classes é fixo e os resultados poderão ser: dados que não estarão propriamente interligados com outras fontes e conseqüentemente não poderão ser reaproveitados; ou a publicação de informações indesejadas, privadas.

Outro aspecto importante diz respeito à abrangência das linguagens de mapeamento. Apesar do enorme número, nenhuma é ampla o suficiente para expressar o mapeamento; não apenas de Banco de Dados Relacionais, como formatos proprietários e de outras fontes encontradas em sites governamentais, como planilhas; soma-se a isso o a dificuldade que alguns usuários encontravam para realizar a manutenção destes mapeamentos, uma vez que tanto o esquema quanto os vocabulários da linguagem de mapeamento e da fonte dos dados pode sofrer alterações ao longo do tempo. R2RML, a linguagem padrão para representar RDB, por exemplo, é constituída de 15 classes, várias propriedades, e passou por três versões em menos de um ano, o que leva muitos usuários a optarem pelo mapeamento direto.

Recentemente a W3C estabeleceu o padrão RDFa [24], que permite instanciar declarações RDF através de anotações de documentos HTML. Desde

então, surgiram várias máquinas de busca [25] capazes de indexar páginas em RDFa. Apesar do aparecimento de várias ferramentas para conversão de dados em RDF, pouco ou nada tem sido feito para garantir o apoio para RDFa. Com R2RML, por exemplo, é possível realizar o mapeamento de fontes de dados relacionais para RDF/XML e N3, mas o que fazer com os demais formatos existentes na Web Semântica como RDFa e JSON?

A necessidade de oferecer suporte a outros formatos além de RDB – *Relational Databases*, aliada à dificuldade de usar e estender as ferramentas existentes, causou o aparecimento de um número significativo de soluções voltadas, sobretudo, à solução de problemas específicos [17]. Além disso, essas soluções têm se mostrado um tanto ineficazes, ou até mesmo inacessíveis para a maioria dos usuários, pois apresentam muitas questões a serem resolvidas, tais como:

- Extensibilidade: muitas ferramentas são criadas para atender a uma funcionalidade específica, demonstrando ser uma verdadeira caixa preta aos usuários que desejam customizá-las ou estendê-las. Em alguns casos, o usuário é obrigado a modificar o código fonte enfrentando problemas de linguagem, suporte e documentação;
- Escalabilidade: algumas ferramentas não garantem robustez na hora de operar sob uma grande quantidade de informações, outras ferramentas não apresentam um desempenho adequado para aplicações profissionais;
- Facilidade de uso: um grande obstáculo para os usuários de maneira geral é a dificuldade de configurar ou operar essas ferramentas. Muitas vezes é preciso aprender linguagens ou, utilizar ferramentas adicionais para atingir o resultado esperado.

É preciso difundir a Web Semântica e fomentar o surgimento de ferramentas mais amigáveis e extensíveis, voltadas para não-especialistas, possibilitando a imersão desses profissionais nesse novo universo para promover um crescimento considerável dos documentos publicados nesse formato.

1.2 Objetivo

Neste trabalho nós propomos Babel, um framework extensível capaz de manipular dados provenientes de várias fontes em um só vocabulário em quaisquer dos formatos da Web Semântica (RDFa, RDF/NTriples, RDF/NQuad ou RDF/XML) utilizando templates.

1.3 Contribuições

As principais contribuições desta dissertação são:

- Um framework extensível que permite a transformação e unificação de informações de várias fontes de dados no formato RDF (RDFa, XML, NTriples ou NQuad);
- TML, uma linguagem de mapeamento baseada em templates que promove a reutilização de vocabulário, possibilitando a usuários familiarizados com as estruturas de dados RDF criar e gerenciar mapeamentos, utilizando modelos produzidos e publicados por outros usuários. TML também pode ser utilizada na modelagem de diferentes bases: relacionais ou formatos de arquivos proprietários;
- Babel GUI, uma interface gráfica que facilita a seleção bem como a conversão e a publicação de informações em bancos de dados RDF [26].

1.4 Organização do trabalho

A dissertação foi estruturada da seguinte forma: no próximo capítulo, faremos uma revisão sobre os fundamentos básicos do framework proposto. Para aqueles familiarizados com a Web Semântica, é sugerido ir direto ao capítulo 4. No capítulo 3 discutimos os trabalhos relacionados. No capítulo 4 apresentamos uma nova linguagem de tempaltes, a TML. O framework é apresentado no capítulo 5. Por fim, no capítulo 6, é apresentada a conclusão e trabalhos futuros.