ARTIGO

Recebido em: 18/12/2017

Aceito em: 08/05/2018

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 23, n. 53, p. 137-146, set./dez., 2018. ISSN 1518-2924. DOI: 10.5007/1518-2924.2018v23n53p137

O papel da web semântica nos processos do *big data*

The role of semantic web in the big data process

Caio Saraiva CONEGLIAN (caio.coneglian@gmail.com) *
Rodrigo DIEGER (rdieger@gmail.com) **
José Eduardo SANTAREM SEGUNDO (santarem@usp.br) ***
Miriam CAPRETZ (mcapretz@uwo.ca)****

* Doutorando e Mestre em Ciência da Informação na Universidade Estadual Paulista (Unesp)
** Mestrando em Ciência da Informação (PPGCI) na Universidade Estadual Paulista (Unesp)
*** Professor da Universidade de São Paulo e professor do Programa de Pós-Graduação em Ciência da Informação (PPGCI) na Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp)
**** Professora do *Department of Electrical and Computer Engineering* na Western University (Canadá)

Resumo

A Web Semântica apresenta um corpus teórico e diversas tecnologias e aplicações que demonstram a sua consistência, inclusive no que tange ao uso de seus conceitos e de suas tecnologias em outros escopos não se limitando unicamente a Web. Neste sentido, os projetos de *Big Data* podem tirar proveito da aplicação dos princípios e dos desenvolvimentos realizados na área da Web Semântica, para aperfeiçoar os processos de análises de dados, em especial na inserção de características semânticas para contextualização dos dados. Assim, esta pesquisa tem como objetivo analisar e discutir o potencial das tecnologias da Web Semântica como meio de integração e desenvolvimento de aplicações de *Big Data*. Utilizou-se uma metodologia qualitativa exploratória, em que se buscaram pontos de convergência entre a Web Semântica e *Big Data*. Foram identificados e discutidos quatro pontos principais: a aplicação do *Linked Data* enquanto fonte de dados para o *Big Data*; o uso de ontologias nas análises de dados; o uso das tecnologias da Web Semântica para promoção da interoperabilidade em cenários de *Big Data*; e o uso de *machine learning* para extrair dados automaticamente e convertê-los para os padrões da Web Semântica. Neste sentido, foi possível identificar que a Web Semântica, em especial no que permeia suas tecnologias e aplicações, pode auxiliar significativamente o desenvolvimento do *Big Data*, por fornecer um paradigma complementar dos aplicados majoritariamente nas análises de dados.

Palavras-chave: Web Semântica. *Big Data*. Tecnologias da Web Semântica.

Abstract

The Semantic Web presents a theoretical corpus and a range of technologies and applications that demonstrate its consistency, including in use of its concepts and its technologies in other scopes than the Web. In this sense, Big Data's projects can take advantage of the application of principles and developments in the area of the Semantic Web, to improve the processes of data analysis, especially in the insertion of semantic characteristics for data contextualization. Thus, this research aims to analyze and discuss the potential of Semantic Web technologies as a means of integrating and developing Big Data applications. An exploratory qualitative methodology was used, where we searched for points of the literature and documentary texts dealt with the convergence between the Semantic Web and the Big Data. Four main points were identified and discussed: the application of Linked Data as a data source for Big Data; the use of ontologies in data analysis; the use of Semantic Web technologies to promote interoperability in Big Data scenarios; and the use of machine learning to extract data automatically and convert them to Semantic Web standards. Therefore, it was possible to identify that the Semantic Web, especially with regard to its technologies, can help Big Data, since it provides a paradigm different from those applied mainly in data analysis.

Keywords: Semantic Web. Big Data. Semantic Web Technologies.

v. 23, n. 53, 2018.
p. 137-146
ISSN 1518-2924

1 INTRODUÇÃO

Vive-se a era do *Big Data*. O intenso processo de evolução e utilização das tecnologias computacionais e informacionais que se tem experimentado nos últimos anos, vem acelerando de maneira radical a expansão e integração dos mais variados dispositivos e ambientes informacionais digitais, impactando a forma como estão sendo criados e utilizados os dados e as informações oriundas destes contextos.

A geração e consumo de dados vem se tornando uma parte importante da vida diária de pessoas e das organizações em geral, particularmente com a disponibilidade e o uso massificados da tecnologia e de aplicações da Internet. Zikopoulos, Eaton e Deroos (2012) definem que a era do *Big Data* é resultado das mudanças que tem ocorrido no mundo, onde por meio dos avanços das tecnologias, foi possível que pessoas e programas se intercomunicassem durante todo o tempo.

Em decorrência deste novo paradigma, observa-se um aumento exponencial no volume, na variedade (fontes, formatos e esquemas distintos) e na velocidade com que dados e informações vem sendo criados e disponibilizados. Estudo publicado pelo *International Data Corp* (IDC) prevê que a criação de dados aumentará para cerca de 163 zettabytes (ZB) até 2025, um aumento de dez vezes nos valores de 2016, e também considera que a coleta, gerenciamento e análise de dados sejam a força motriz por trás de quase todas as atividades humanas na próxima década (GANTZ; REINSEL, 2017).

Esse rápido e contínuo crescimento, somado às limitações dos métodos e formas tradicionais de análise e de processamento (levando-se em conta suas características), apresentam inúmeros desafios relacionados à maneira como tornar estes dados e informações disponíveis para uso de maneira efetiva. Beyer e Laney (2012) definem *Big Data* como o alto volume, alta velocidade e/ou alta variedade de informações que requerem novas formas de processamento para permitir melhor tomada de decisão, nova descoberta do conhecimento e otimização de processos.

Apesar de provenientes de uma direção diferente ao *Big Data*, os conceitos e as tecnologias da Web Semântica permitem reunir fontes heterogêneas de dados para explorar e fornecer significado a diferentes conjuntos, facilitando a aplicação do processamento semântico. A partir da interoperabilidade de tecnologias e conceitos desses diferentes campos, permite-se um novo processo de descoberta de conhecimento, agrupando e organizando a informação disponível de maneira eficiente e integrada, permitindo dessa forma que se explore, analise, processe e transforme dados a partir de fontes distintas.

Diante deste cenário, o objetivo deste artigo é analisar e discutir o potencial das tecnologias da Web Semântica como meio de integração e desenvolvimento de aplicações de *Big Data*. Além disso, procura demonstrar os principais desafios da integração de dados relacionados com este tema. Para atingir os objetivos deste trabalho, utilizou-se uma metodologia qualitativa exploratória, em que se buscaram pontos em que a literatura e textos documentais tratavam da convergência entre as tecnologias e os conceitos da Web Semântica e os processos que tangenciam o *Big Data*. Para realizar a pesquisa, identificou-se primeiramente temáticas de estudos em que há essa relação iminente da aplicação das tecnologias da Web Semântica no cenário do *Big Data*. Posteriormente, foi realizado uma explanação sobre cada um dos pontos identificados, apontando como ocorre o uso das tecnologias da Web Semântica, além de verificar como esta utilização contribui para os processos de *Big Data* como um todo.

O texto foi organizado com uma introdução, seguido das seções tratando dos pressupostos teóricos de *Big Data* e de Web Semântica. Em seguida, são apresentados os resultados e discussões e as considerações finais.

2 *BIG DATA*

Nos últimos anos, podemos observar de maneira significativa o avanço exponencial no número de pesquisas e aplicações que desenvolvem e exploram os conceitos relacionados a *Big Data*. Laney (2001), em uma das primeiras definições sobre este tema, afirma que o *Big Data* se caracteriza essencialmente a partir de três aspectos: volume, velocidade e variedade.

No que tange ao volume, há uma relação direta com o tamanho e a quantidade de dados. Neste quesito, nos tempos atuais a geração de dados cresce de forma exponencial, em que em poucos anos gera-se uma quantidade de dados imensurável. Taurion (2013) relata

que atualmente, grande parte dos dados se encontram disponíveis em formatos digitais. Esta migração do analógico para o digital pode ser apontada como uma das principais responsáveis pela imersão deste cenário de *Big Data*.

A velocidade refere-se a aspectos da dinâmica de crescimento e processamento dos dados, em que o sucesso de uma tomada de decisão está muito interligado ao momento e a rapidez que tal decisão é tomada. Outro ponto é relativo ao acesso em tempo real das informações, o aspecto da velocidade também está interligado a necessidade de por vezes os gestores terem os dados em tempo real, algo que foi possibilitado principalmente pelas novas tecnologias de informação e comunicação.

O terceiro aspecto, a variedade, trata principalmente à diversidade de origens, formas e formatos dos dados. A variedade se configura como o elemento mais desafiador do *Big Data*, pois os ferramentais que tratam com dados no contexto do *Big Data* devem verificar e trabalhar com dados estruturados, semiestruturados e não estruturados, o que exige dos mecanismos computacionais flexibilidade e adequabilidade para extrair o valor de dados tão diversos. (DEMCHENKO et al., 2013).

Com o passar dos anos, a compreensão acerca do *Big Data* e como esse fenômeno está alterando a nossa sociedade está aumentando. Mayer-Schönberger e Cukier (2013) no livro "*Big Data*: Como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana", aprofundam a discussão da temática do *Big Data*, enfocando-se bastante nas questões relativas as análises de dados em cenário de *Big Data*.

Os autores ao tratar desses pontos afirmam que:

> Antes do big data, nossa análise geralmente se limitava a uma pequena quantidade de hipóteses que definíamos bem antes de coletarmos os dados. Quando deixamos que os dados falem por si, podemos gerar conexões que nem sabíamos que existiam. Assim, alguns fundos hedge [forma de investimento alternativa] usam o Twitter para prever o desempenho do mercado de ações. A Amazon e a Netflix baseiam suas recomendações de produtos nas diversas interações em seus sites. Twitter, LinkedIn e Facebook mapeiam o 'gráfico social' das relações entre os usuários para aprender mais sobre suas preferências. (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 9)

A partir da afirmação dos autores, verifica-se que há uma mudança estrutural na forma como se realiza análises, que consequentemente afeta a tomada de decisão dos gestores. Neste âmbito, é necessário que os dados sejam analisados como um todo e na sua totalidade, não se limitando a dados amostrais ou tampouco partindo de algumas hipóteses pré-definidas.

Os cenários apontados sobre a análise de dados conduzem a diversas discussões sobre as técnicas e métodos que devem ser utilizados para a realização de análises de dados. Em um dos estudos que buscam enumerar e direcionar as análises de dados, Bugembe (2016) aponta as fases necessárias para extrair valor dos dados. O autor afirma que as seis fases são: 1) fonte, 2) captura e armazenamento, 3) processamento e fusão, 4) acesso, 5) análise e 6) exposição.

Outro ponto importante está na interação com as mídias sociais e a Web, que passou a ser um elemento importante para que os dados gerados nestes ambientes proporcionassem análises acerca dos comportamentos e de padrões dos consumidores e de potenciais consumidores.

Neste contexto, a propagação e disseminação de dados oriundos das redes sociais, comunicação entre máquinas, sensores, bem como a análise e aproveitamento de artefatos digitais e bases de dados existentes, ou ainda tecnologias emergentes como a "Internet das Coisas" e o fenômeno dos dados abertos, produzem-se em larga escala e tornam praticamente qualquer coisa como dado ou conteúdo, que precisam ser cada vez mais bem interpretados e examinados. No entanto, a maioria desses dados é ainda inacessível, pois precisamos de tecnologia e ferramentas para encontrar, transformar, analisar e visualizar dados para torná-los consumíveis para a tomada de decisões (BANSAL, 2014).

Assim, o uso de tecnologias que buscam aprofundar a compreensão do significado e do contexto dos dados pode auxiliar para melhorar os processos analíticos do *Big Data*. As tecnologias da Web Semântica são atualmente os principais instrumentos computacionais

que estão conseguindo realizar tal função. Desta forma, na próxima seção apresenta-se os conceitos teóricos da Web Semântica, além de apontar as suas principais tecnologias.

3 WEB SEMÂNTICA

A crescente geração e disponibilização de dados está diretamente vinculada a evolução das Tecnologias de Informação e Comunicação, que foram se popularizando com o passar dos anos e se tornaram essenciais no cotidiano dos indivíduos. Neste sentido, a reflexão acerca do *Big Data* na dimensão atual necessita que as tecnologias sejam disseminadas e utilizadas em larga escalar pelas pessoas, pois o valor dos dados está muitas vezes em obter padrões de comportamentos humanos.

Diante desta seara, a Web se tornou uma das principais tecnologias de uso cotidiano das pessoas. Desde tarefas básicas como mandar um e-mail ou assistir um filme, até a interação do indivíduo com o governo, passou a se dar neste ambiente. A navegação no ciberespaço passou assim, a ser constante na vida das pessoas, em que a Web é o principal ambiente e o proporcionador desta disrupção que a sociedade vive nos tempos atuais.

A Web foi concebida em 1989 por Tim Berners-Lee, sendo uma proposta que utilizava o conceito do hipertexto para disponibilizar informações utilizando a infraestrutura da Internet. Rapidamente a Web foi crescendo de forma descentralizada, se popularizando entre os mais diversos tipos de usuários. A consequência desse crescimento, foi o aumento dos dados disponibilizados neste ambiente, que estavam em sua maioria estruturados somente para a leitura humana, não havendo esquemas que propiciassem a mecanismos computacionais compreenderem o sentido destas informações.

Uma implicação desse processo foi tornar árdua a tarefa dos usuários recuperarem conteúdo da Web, além de dificultar que agentes computacionais fossem capazes de explorarem e realizarem inferências acerca dos dados disponibilizados. Ainda que os motores de busca tenham aperfeiçoado o processo de localização de um conteúdo na Web ao longo dos últimos anos, o nível de semântica formal neste ambiente era bastante raso, sem possibilitar com que a descoberta de informações fosse eficiente.

Visando solucionar tal conjuntura, Berners-Lee, Hendler e Lassila propuseram em 2001 a Web Semântica. Estes autores relataram que a Web Semântica seria uma Web na qual os computadores poderiam entender o contexto das pessoas, para poder interpretar o significado da informação (BERNERS-LEE; HENDLER; LASSILA, 2001).

Os autores afirmam ainda que: “A Web Semântica não é uma Web separada, mas uma extensão da atual, em que a informação tem um significado bem definido, permitindo que os computadores e as pessoas trabalhem em cooperação.” (BERNERS-LEE; HENDLER; LASSILA, 2001, não paginado, tradução nossa).

Com a evolução da Web Semântica, uma gama de tecnologias foi desenvolvida visando tornar materializável a proposta. Dentre as tecnologias, destaca-se o *Resource Description Framework* (RDF) como o modelo padrão para intercâmbio de dados na Web. O RDF tem a característica de estruturar os dados nas chamadas tripas, que interligam um recurso a outro ou a um valor por meio de uma relação.

Outra tecnologia de fundamental importância para a Web Semântica é a ontologia. Uma ontologia busca descrever computacionalmente e formalmente um determinado domínio. Essa tecnologia é essencial para que um agente computacional seja capaz de compreender a semântica formal dos recursos, e assim, permitir a realização de inferências. A *Web Ontology Language* (OWL) é a linguagem recomendada pela *World Wide Consortium* (W3C) para a construção de ontologias, utilizando o conceito do RDF para interligar os recursos.

Por fim, destaca-se o *SPARQL Protocol and RDF Query Language* (SPARQL) como o protocolo para a realização de consultas dos dados que seguem os princípios da Web Semântica. Tal protocolo utiliza como base as triplas RDF para a realização das consultas, o que a torna fundamental para os mecanismos computacionais recuperarem informações na Web.

A partir do desenvolvimento e do estabelecimento das tecnologias da Web Semântica, diversas aplicações foram desenvolvidas, em um processo chamado de materialização da Web Semântica (SANTAREM SEGUNDO; CONEGLIAN, 2016). Dentre as aplicações e as iniciativas criadas, o *Linked Data* possui maior destaque, por ser amplamente difundida e utilizada em diversas áreas do conhecimento.

O *Linked Data* tem como base os conceitos e as tecnologias da Web Semântica, bem como uma série de padrões e boas práticas, para ser uma diretriz para a publicação de dados na Web. Nesta iniciativa, os dados possuem um valor semântico explícito, contendo interligações com outras bases de dados e permitindo inferências por ferramentas computacionais.

A partir da proposta do *Linked Data*, diversas instituições começaram a disponibilizar seus dados abertamente seguindo essa proposta, possibilitando hoje que as aplicações possam usar esses dados para gerar valor. Domínios como da saúde e de publicações estão engajados nessa proposta, disponibilizando grandes quantidades de dados e de *datasets* abertamente.

A partir das tecnologias da Web Semântica e das aplicações construídas com base nelas, há a possibilidade de realizar a sua integração com o *Big Data*. A busca de inserir e aprimorar os processos analíticos do *Big Data* com a Web Semântica será discutida a seguir na seção de resultados e discussões.

4 RESULTADOS E DISCUSSÕES

A partir dos pressupostos teóricos apontados, identifica-se cenários em que a aplicação da Web Semântica pode ocorrer no âmbito do *Big Data*. Passando desde os pontos relativos às próprias fontes de informações, até na inserção de um número maior de argumentos nas análises de dados, a Web Semântica, juntamente com alguns de seus conceitos, tecnologias e aplicações pode trazer semântica e contextualização nos processos que se relacionam ao *Big Data*.

Neste sentido, questões que permeiam o significado dos dados desempenham um papel fundamental no que se refere ao uso efetivo e ao aproveitamento das informações e do conhecimento extraídos. Para enfrentar esses desafios, as tecnologias e os conceitos de diferentes campos podem ser combinados, permitindo um avançado processo de descoberta de conhecimento.

Quando direcionamos nossa abordagem para o significado dos dados, os conceitos e as tecnologias da Web Semântica se apresentam de maneira proeminente e definem um componente estratégico para a tratativa da variedade de dados no cenário do *Big Data*.

Esses padrões semânticos possuem recursos compatíveis com as necessidades de dados existentes e estrito alinhamento com o *Big Data*. Características que de maneira geral refletem sobre representação do conhecimento, interoperabilidade de dados, e recuperação da informação também definem um importante aspecto neste contexto para resolver questões relacionadas com análise e a variedade de dados.

No âmbito deste trabalho, enfoca-se no aspecto da variedade (heterogeneidade de dados e diversidade das fontes de dados) pela diversidade de se tratar isto no âmbito computacional. Para tratar a heterogeneidade, a abordagem semântica é a que melhor se apresenta para resolver estas problemáticas. Para entender, relacionar e interpretar dados, é necessário o significado explícito dos dados, que é dado pelo aproveitamento efetivo das tecnologias e abordagens semânticas.

Ao analisar diversas literaturas sobre tecnologias da Web Semântica e *Big Data*, identifica-se que estas desempenham um papel importante para converter dados em conhecimento. Em comparação com outras tecnologias, as tecnologias semânticas fornecem conhecimento prévio para o contexto dos dados, interoperabilidade, escalabilidade, integração e aceitos como padrão de expressividade de dados.

Neste contexto, apresenta-se na sequência os principais pontos em que a Web Semântica pode denotar um papel estratégico e de grande relevância principalmente para a tratativa da variedade e a descoberta de novas relações e padrões entre os grandes volumes de dados que se apresentam em um cenário de *Big Data*.

4.1 *Linked Data*: conectando o *Big Data*

O meio como as informações estão estruturadas em cenários de *Big Data* é significativamente distinto daqueles conjuntos de dados estruturados seguindo os princípios do *Linked Data*. Em suma, a maioria dos dados tratados como *Big Data* são desestruturados ou semiestruturados, enquanto na perspectiva do *Linked Data,* são integralmente estruturados.

A diferença entre estes dois cenários é acentuada pela existência de metadados que apontem o contexto e o significado que os conjuntos de dados estabelecem dentro do *Linked Data*, e que de modo geral não se refletem no contexto do *Big Data*. Desta forma, os dados de *Linked D*ata tornam-se uma importante fonte de informação, ao fornecer dados estruturados e com semântica formal, tratando de um domínio específico.

No entanto, o *Linked Data* contempla um escopo limitado de conjuntos de dados, que foi tratado e enriquecido a partir de procedimentos computacionais em ambientes minimamente controlados, tendo assim, função e princípios diferentes do *Big Data*, que irá contemplar dados das mais variadas fontes, sem apresentar um rígido controle sobre a estrutura destes dados. Assim, o *Linked Data* não pode ser utilizado como um substituto das fontes informacionais de grande volume do *Big Data*, mas sim um elemento complementar nos processos de análises de dados.

Há diversas correntes defendidas, como Bugembe (2016), sobre os métodos utilizados durante os processos de análises de dados, que irão apresentar os pontos que devem ser considerados, bem como as fases aplicadas para a análise. Como relatado anteriormente, um destes autores é Bugembe (2016), que divide em seis o que ele chama de fases para obtenção de valor dos dados durante as análises: 1) fonte; 2) captura e armazenamento; 3) processamento e fusão, 4) acesso; 5) análise; e 6) exposição.

O autor, ao discutir essas diversas fontes, vai inserindo fase a fase como deve ser realizada a coleta dos dados, as preocupações quanto a escolha das fontes, o processamento, a análise, entre outros. Desta forma, identifica-se sempre a busca por relacionar informações relevantes e que possam de alguma forma possuir confiabilidade. Neste sentido, o *Linked Data* se mostra como uma fonte auxiliar aos dados, capaz de fornecer aos processos subsequentes uma maior confiabilidade, além de permitir que as relações realizadas nos processos de fusão, ocorram com um número maior de argumentos, permitindo ainda que fontes relacionadas sejam incluídas e utilizadas durante o processo.

Em síntese, o *Linked Data* traz dados estruturados e semanticamente formalizados ao processo de análise, permitindo com que a exploração dos dados brutos (não estruturados e semiestruturados) na busca de extrair *insights* e padrões comportamentais, seja aprimorado ao considerar uma fonte que permita contextualizar e conduzir a realização de inferências com um nível lógico mais profundo nesta integração entre o *Linked Data* e os demais dados. Um instrumento que contribui para o *Linked Data* e que pode aprimorar nos processos de *Big Data* são as ontologias, exploradas na sequência.

4.2 Ontologias como estratégia para a análise e organização do conhecimento

As ontologias são instrumentos centrais para a Web Semântica por representarem formalmente um determinado domínio, explicitando axiomas nas relações existentes entre os recursos. Essa característica discutida por Santarem Segundo e Coneglian (2016), demonstra o potencial computacional que as ontologias possuem ao representar um determinado domínio, promovendo a realização de inferências quando se usa as ontologias na descoberta de informações.

Desta forma, o uso de ontologias pode ocorrer em diversas etapas das análises de dados em cenários de *Big Data*, por possibilitar um nível de semântica formal essencial nos processos que visam extrair valor dos dados.

Um possível uso das ontologias neste contexto, se caracteriza pela necessidade de pesquisadores da área de *Big Data* explorarem o poder das correlações estatísticas ao analisar grandes conjuntos de dados que podem estar relacionados, e assim extrair algum valor destas massas de dados. Mayer-Schönberger e Cukier (2013) afirmam que: "Previsões com base em correlações estão na essência do *Big Data*", o que demonstra como as teorias lógicas, matemáticas e estatísticas auxiliam significativamente na tomada de decisão dos gestores ao analisar os dados.

Neste sentido, as ontologias por serem um aparato tecnológico capaz de expressar um domínio com lógicas, e com capacidade representacional que permite a realização de inferências, podem trazer um suporte significativo nestes processos que estão inter-relacionando bases de dados, e assim permitindo a realização de predições.

Pereira Junior et al. (2016, p. 103, tradução nossa) discorre sobre a possibilidade do uso de ontologias para a fusão de informação, afirmando que os processos tradicionais de fusão são baseados unicamente na sintaxe, ao invés do significado dos termos, enquanto a

fusão semântica com ontologias "[...] permite gerar informações com qualidade aprimorada e mais fiel ao ambiente real."

Diante desses pontos, o uso das ontologias na fusão de dados surge como um meio de tornar os resultados desse processo computacional mais aprimorado e eficiente, trazendo ao *Big Data* a inserção da semântica e do contexto na análise em si. Tal questão se mostra como um contraponto aos métodos de análises que se focam unicamente nas relações estatísticas e matemáticas dos dados, que não deixam de ter valor, mas passam a ser complementadas por uma análise mais profunda do contexto que os dados se encontram.

Uma consequência da adoção de ontologias para a realização das chamadas fusões de informações semânticas, seria a possibilidade de tornar o processo de análise, discutido por Bugembe (2016) mais aprimorado, por ter um instrumento informacional que embasa a realização da fusão e possibilita inferências nesta fase de análise, a partir dos axiomas das propriedades das ontologias. Outro ponto promovido pelas ontologias trata da interoperabilidade, que se mostra como um outro ponto essencial para o *Big Data* e que pode ser aprimorado a partir dos conceitos e das tecnologias da Web Semântica.

4.3 *Big Data* e os desafios da interoperabilidade semântica dos dados

Interoperabilidade de dados pode ser contextualizada a partir da capacidade fornecida aos sistemas para interpretar de maneira automática e precisa o significado dos dados trocados. Para alcançar a interoperabilidade de dados semânticos, os sistemas não precisam apenas trocar seus dados, mas também trocar ou concordar com modelos explícitos desses dados (HARMELEN, 2008).

No contexto de *Big Data*, dados oriundos de fontes não estruturadas e heterogêneas se estabelecem como uma de suas principais características. Alcançar a interoperabilidade semântica nestes casos pode ser considerado um grande problema, visto principalmente a variedade de características e particularidades de cada fonte de dados observadas a partir deste cenário.

As tecnologias e os conceitos da Web Semântica permitem aplicar enriquecimento semântico aos dados por meio do uso de vocabulários específicos, ontologias e padrões de metadados. Além disso, outra vantagem apresentada por este modelo fundamenta-se no fato de ser um padrão estabelecido para que os dados sejam lidos e interpretados a partir de agentes computacionais, promovendo uma autonomia e independência para os sistemas que fazem uso efetivo dos dados concebidos a partir deste modelo, permitindo reduzir o custo e a complexidade da integração de dados.

As soluções atuais de processamento, armazenamento e recuperação de dados heterogêneos e distribuídos no contexto do *Big Data*, oferecem níveis de escalabilidade, robustez, tolerância a falhas e elasticidade sem precedentes. No entanto, não é possível compartilhar o potencial das tecnologias da Web Semântica em grande parte dessas soluções, visto que os valores atribuídos aos dados normalmente não possuem uma anotação semântica explícita. Isto ocorre pois os dados Assim, a possibilidade de combinar dados não estruturados em grande escala com dados estruturados e tecnologias da Web Semântica, expande as oportunidades em *Big Data* de processar dados de novas formas e combinações.

Victorino et al. (2017) aponta uma proposta de um ecossistema de *Big Data* para análises de dados abertos governamentais, em que tecnologias da Web Semântica, como ontologias, dão suporte a realização de interoperabilidade e de processos analíticos dos dados abertos. Este trabalho demonstra como as tecnologias da Web Semântica podem contribuir efetivamente, estando integrado com as principais ferramentas de *Big Data* existentes.

Padrões da Web Semântica e *Linked Data* como o RDF (*Resource Description Framework*), que conforme define a W3C (2004), tem como um dos principais objetivos criar uma rede de informações a partir de dados distribuídos, e o protocolo SPARQL (*Simple Protocol and RDF Query Language*) para recuperação da informação em ambientes semânticos, destacam-se como exemplos concretos na direção de oportunidades e alternativas estratégicas para a problemática da interoperabilidade de dados semânticos na era do *Big Data*.

4.4 *Machine Learning*: Extrair e estruturar dados com o auxílio das Tecnologias da Web Semântica

Outro ponto importante no que tange às análises de *Big Data* está na coleta automatizada de dados que estão disponíveis abertamente. Neste contexto, diversos estudos apontam que quando se trata de tomada de decisão, grande parte dos dados necessários encontram-se disponíveis abertamente na Web.

Um dos problemas de extrair esses dados está na falta de estrutura ou na falta de padronização das estruturas utilizadas. Assim, muitas vezes é necessário que sejam desenvolvidos diversos robôs extratores, um para cada ambiente ou tipo de estrutura, dificultando o processo de coleta dos dados.

Para o *Big Data*, essa configuração apontada não é vantajosa, pois a busca é justamente em extrair e obter a maior quantidade possível de dados, para que as análises realizadas sejam capazes de considerar um maior nível de variáveis, e a tomada de decisão seja a mais embasada possível.

Diante deste cenário, o uso de *machine learning* pode auxiliar com que o processo de extração de dados ocorra com mais facilidade, para que não haja a necessidade de construir diversos extratores distintos. O uso do *machine learning* seria justamente na busca de tornar o processo de coleta mais automático, em que por meio das técnicas de Inteligência Artificial, como o treinamento e o reconhecimento de padrões, um agente computacional extrairia as informações necessárias de um documento ou de uma página não estruturada, como por exemplo um arquivo de texto PDF.

Com o uso de *machine learning* para reconhecer e extrair dados, seria possível estruturar as informações seguindo os princípios da Web Semântica, podendo criar *datasets* em RDF em que as relações existentes são explicitadas e podem conter um nível de semântica formal adequado. A ideia central estaria em coletar dados não estruturados na Web, como em arquivos PDF, utilizando técnicas de *machine learnin*g para identificar o que deve ser extraído, sendo realizado posteriormente uma conversão e uma estruturação dos dados para os formatos da Web Semântica, como o RDF.

Desta forma, os dados seriam coletados de diversos ambientes e seriam uma fonte para as análises de dados em *Big Data* enriquecida semanticamente e que poderia ser consultada por meio das tecnologias da Web Semântica, como o SPARQL. Como relatado anteriormente, o uso de dados seguindo os princípios do *Linked Data*, seria importante para aprimorar os processos do *Big Data*, ao fornecer dados que permite a realização de inferências e sendo uma fonte estruturada em meio de tantos dados não estruturados ou semiestruturados.

Uma ferramenta que pode auxiliar neste processo é o *GeneRation Of BIbliographic Data* (GROBID), que é uma biblioteca de programação de *machine learning* que busca extrair, analisar e estruturar dados brutos, como documentos em PDF. A ferramenta arranja os dados no formato *Text Encoding Initiative* (TEI), uma estrutura para organizar textos.

O GROBID busca identificar informações de um arquivo PDF, como título e resumo, além de ser capaz de verificar questões relativas a patentes, identificar nomes, realizar análises em datas, entre outros. Tal ferramenta utiliza as técnicas de *machine learning* para aprimorar estes processos e identificar diferentes estruturas de documentos.

Um outro exemplo de ferramenta que auxilia no uso de *machine learning* é o "scikit-learn". Uma biblioteca para a linguagem de programação *python*, que auxilia em análises de dados e mineração de dados. O uso de bibliotecas como a citada pode ser um instrumento para obter informações relevantes de grandes massas de dados e assim, posteriormente serem estruturados utilizando as tecnologias da Web Semântica.

Há diversas outras bibliotecas e ferramentas de *machine learning* que podem auxiliar nesse processo de extração de dados, auxiliando significativamente na coleta automatizada dos dados. Consequentemente, a partir dos dados coletados, a transformação nos formatos da Web Semântica pode ocorrer com tecnologias que a partir de dados estruturados, os converte em RDF.

4.5 Síntese do papel da Web Semântica nos processos de Big Data

A Web Semântica e as suas tecnologias estão ligadas em diversos momentos dos processos de análises de dados em cenários de *Big Data*. Há diversas tecnologias que podem auxiliar nesse processo, como foi discutido anteriormente. Em síntese, pode-se dividir em três momentos que as tecnologias da Web Semântica podem auxiliar no aprimoramento dos processos do *Big Data*: fontes, análises e resultados.

Nas fontes, a Web Semântica está principalmente em dois pontos: 1) o *Linked Data* enquanto uma fonte de informação estruturada com um nível de semântica formal elevado, sendo utilizado para fornecer mais um elemento para a realização das análises; e 2) a transformação dos dados não estruturados ou semiestruturados em RDF, com o auxílio de agentes utilizando *machine learning*, em que os agentes irão extrair e coletar os conteúdos espalhados pela Web, realizando posteriormente uma estruturação seguindo os princípios da Web Semântica.

Os dois pontos apontados estão focados principalmente em fornecer fontes de dados mais significativas, em que seja possível explorar com mais êxito a realização de inferências, tendo o significado dos conteúdos. Consequentemente, ao utilizar tais fontes, juntamente com as tradicionais fontes de *Big Data*, será possível obter melhores resultados, uma vez que será possível relacionar e aprofundar as análises nas grandes massas de dados.

O segundo momento apontado, está na análise de *Big Data*, que a Web Semântica pode contribuir principalmente fornecendo as ontologias. Neste caso, as ontologias serão um elemento fundamental para tornar as análises realizadas mais contextualizadas, visto que fornecem aos instrumentos computacionais, uma visão acerca do domínio em que está sendo analisado, que é capaz de ser compreensível computacionalmente. Desta forma, as ontologias poderão ser utilizadas para aprimorar as correlações estatísticas das análises realizadas, bem como um instrumento para tornar a fusão de informações semanticamente mais enriquecidas.

Por fim, no que tange aos resultados obtidos a partir das análises de dados em cenários de *Big Data*, foi apresentado sobre a interoperabilidade semântica dos dados, em que há um desafio de permitir a troca e a recuperação dos resultados obtidos. As tecnologias como RDF e SPARQL podem auxiliar nisso, ao permitir com que os resultados sejam estruturados e utilizados em outras oportunidades.

Os três momentos apontados demonstram que as tecnologias da Web Semântica podem influenciar as análises de uma ponta a outra do processo, ou seja, desde a coleta dos dados, passando pela análise, chegando até os resultados. Vale destacar que as análises de *Big Data* não são contempladas unicamente pelas tecnologias da Web Semântica, mas podem ser aprimoradas por meio delas.

5 CONSIDERAÇÕES FINAIS

A Web Semântica a partir da sua concepção original em 2001, vem evoluindo significativamente, em especial no que tangencia a criação de conceitos e de tecnologias que possam promover os princípios idealizados por seus criadores. Diante dessa evolução, a Web Semântica transcendeu as barreiras da própria Web, fornecendo instrumentos que auxiliam ferramentas computacionais nos mais diversos âmbitos, inclusive em bases de dados privadas e corporativas. Isso se estabeleceu principalmente pela forma como a Web Semântica passou a conceber o tratamento dos dados, contribuindo com instrumentos que favorecem a contextualização em um determinado domínio.

Um cenário que se apresentou como expoente na utilização das contribuições da Web Semântica ao fornecer meios para a realização de inferências e de lógicas e processos para descoberta de conhecimento, foi o *Big Data*, em especial para a realização de análises de dados que se enquadram neste contexto. Essa união entre os processos do *Big Data* com as tecnologias da Web Semântica pode ser estratégica e fundamental para tornar as análises mais efetivas, considerando um número maior de argumentos, a partir de fontes organizadas e estruturadas, apresentando uma maior contextualização do domínio que está sendo analisado. Diante de tais pontos, esta pesquisa buscou identificar e apresentar algumas intersecções existentes entre os processos de *Big Data* e as tecnologias da Web Semântica, indicando como estas últimas contribuíram para aprimorar em especial as análises de dados realizadas.

A utilização do *Linked Data* como fonte de dados, o uso de ontologias para aprimorar os processos de fusão e análises, o aperfeiçoamento da interoperabilidade no *Big Data* e o uso de *machine learning* para extrair dados e estrutura-los com as tecnologias da Web Semântica, foram os quatro pontos que foram discutidos nesta pesquisa, apontando alguns detalhes sobre como se daria a aplicação de algumas tecnologias da Web Semântica para tornar os processos de *Big Data* mais contextualizado semanticamente.

Portanto, esta pesquisa avança na intersecção entre estes dois campos de estudos, discorrendo sobre como a Web Semântica, que apresenta um corpus teórico mais consistente, para tornar o *Link* mais eficiente ao inserir uma ótica semântica nos processos analíticos. Enquanto trabalhos futuros, busca-se realizar a implantação de experimentos que comprovam na prática a viabilidade dos pontos discutidos.

REFERÊNCIAS

BANSAL, S. K. Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration. In: IEEE INTERNATIONAL CONGRESS ON BIG DATA, 2014, Anchorage. **Anais...** p. 522-529. Disponível em: < http://bit.ly/2ulZ3a5>.Acesso em: 22 jul.2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **The Semantic Web**, v. 284, n. 5, p. 28–37, maio 2001.

BERNERS-LEE, T. **Linked Data Principles**. 2006. Disponível em < http://bit.ly/1x6N7XI>. Acesso em: 10 maio 2018.

BEYER, M. A., LANEY, D. **The importance of "Big Data'**: a definition. Stamford, CT: Gartner, 2012.

BUGEMBE, M. **Finding Value in Data**: Determining Where Data Science has The Greatest Impact. O'Reilly: Sebastopol, 2016.

DEMCHENKO, Yuri et al. Addressing big data issues in scientific data infrastructure. In: IEEE INTERNATIONAL CONFERENCE ON COLLABORATION TECHNOLOGIES AND SYSTEMS, 2013, San Diego (USA). **Anais...** p. 48-55.

GANTZ, J., REINSEL, D. **Data Age 2025**: the evolution of data to life-critical: don't focus on big data; focus on the data that's big. 2017 Disponível em: <http://bit.ly/2tPWoU8>. Acesso em: 10 maio 2018.

HARMELEN, F. Semantic Web Technologies As The Foundation For The Information Infrastructure. In: VAN OOSTEROM, P.; ZLATANOVA, S. (Eds.). **Creating Spatial Information Infrastructures**. [S.l.]: CRC Press, 2008. p. 37–52.

LANEY, D. 3D Data Management: Controlling Data Volume, Velocity and Variety. **META group research note**, v. 6, n. 70, 2001.

MAYER-SCHÖNBERGER, V; CUKIER, K. **Big data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. 1. ed. Rio de Janeiro: Elsevier, 2013

PEREIRA JUNIOR, V. A. et al. Using Semantics to Improve Information Fusion and Increase Situational Awareness. In: ADVANCES IN SAFETY MANAGEMENT AND HUMAN FACTORS, Flórida, 2016. **Anais...** Springer International Publishing, 2016. p. 101-113.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. **Inf & Inf**, v. 21, n. 2, p. 217–244, dez. 2016. Disponível em: <http://bit.ly/2uLpbgL>. Acesso em: 10 maio 2018.

TAURION, C. **Big data**. Rio de Janeiro: Brasport, 2013.

VICTORINO, M. C. et al. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais conectados. **Informação & Sociedade**, v. 27, n. 1, 2017. Disponível em: <http://www.periodicos.ufpb.br/index.php/ies/article/download/29299/17505>

W3C. Resource Description Framework (RDF). 2004. Disponível em: <https://www.w3.org/RDF/>. Acesso em: 10 maio 2018.

ZIKOPOULOS, P.; EATON, C.; DEROOS, D. **Understanding BigData**: Analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill, 2012

Editores do artigo: Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.