O PAPEL ESTRATÉGICO DA WEB SEMÂNTICA NO CONTEXTO DO BIG DATA

The Strategic Role of Semantic Web in the Big Data Context

Caio Saraiva Coneglian[1], Rodrigo Dieger[1], José Eduardo Santarém Segundo[2], Miriam Captrez[3]

(1) Universidade Estadual Paulista (UNESP), Av. Hygino Muzzi Filho, 737, Mirante, Marília - SP, 17.525-900, {caio.coneglian, rdieger}@gmail.com,

(2) Universidade de São Paulo (USP), Av. Bandeirantes, 3900 - Vila Monte Alegre, Ribeirão Preto - SP, 14040-900, santarem@usp.br

(3) Western University, 1151 Richmond St, London, ON N6A 3K7, Canadá, mcapretz@uwo.ca

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, onde buscou-se pontos de convergência entre a Web Semântica e Big Data. Foram identificados e discutidos três pontos principais: a aplicação do Linked Data enquanto fonte de dados para o Big Data; o uso de ontologias nas análises de dados; e o uso das tecnologias da Web Semântica para promoção da interoperabilidade em cenários de Big Data. 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. Three 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. 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.

1 Introdução

Vive-se a era do *Big Data*. O intenso processo de evolução e utilização das tecnologias computacionais e informacionais que se tem vivenciado 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 uso massificados da tecnologia e 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 e REINSEL, 2017).

Esse rápido e contínuo crescimento, somado às limitações dos métodos e formas tradicionais de análise e 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.

O texto foi organizado com uma introdução, seguido de uma seção tratando dos pressupostos teóricos tanto de *Big Data* quanto de Web Semântica. Em seguida, são apresentados procedimentos metodológicos do trabalho, finalizando com os resultados e discussões e as considerações finais.

2 *Big Data* e Web Semântica: além das fronteiras da Web

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, variedade. Volume está estritamente relacionado ao tamanho e quantidade de dados. Velocidade refere-se a aspectos da dinâmica de crescimento e processamento dos dados. Variedade à diversidade de origens, formas e formatos dos dados (DEMCHENKO et al., 2013).

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).

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*.

O conceito da Web Semântica foi concebido a partir de 2001, apontando uma Web na qual os computadores poderiam entender o contexto das pessoas, para poder

interpretar o significado da informação (BERNERS-LEE; HENDLER; LASSILA, 2001). As tecnologias da Web Semântica permitem que se criem repositórios de dados, se construam vocabulários e se estabeleçam regras para definição e representação dos dados na Web, mas não se limitando a ela. Além disso, apresenta conceitos e tecnologias para representar conhecimento e suas relações, utilizando uma série padrões e ainda um conjunto de melhores práticas para a publicação de dados estruturados no *Linked Data* (BERNERS-LEE, 2006).

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.

Acredita-se que o desafio técnico mais importante hoje na gestão de *Big Data* é o aspecto da variedade (heterogeneidade de dados e diversidade das fontes de dados). 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 estudar 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.

3 Procedimentos Metodológicos

Para atingir os objetivos deste trabalho, utilizou-se uma metodologia qualitativa exploratória, onde buscou-se 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.

4 Resultados e Discussões

A partir dos procedimentos 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 contexto, na sequência busca-se apresentar 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 semi-estruturados, 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 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. 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* traria 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 aos processos de 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 e 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. 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*.

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 instrumentos 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 e o aperfeiçoamento da interoperabilidade no *Big Data*, foram os três 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 *Big Data* 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. **IEEE**, jun. 2014. 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., 2006. **Linked Data Principles**. Disponível em < http://bit.ly/1x6N7XI>. Acesso em: 15 jul. 2017

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

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: Collaboration Technologies and Systems (CTS), 2013 **International Conference on**. IEEE, 2013. p. 48-55.

GANTZ, J., REINSEL, D., 2017. **Data Age 2025**: The Evolution of Data to Life-Critical. Don't Focus on Big Data; Focus on the Data That's Big. Disponível em: <http://bit.ly/2tPWoU8>. Acesso em: 21 jul. 2017.

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. 2001.

MAYER-SCHÖNBERGER, V; CUKIER, K. **Big data**: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 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. **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**, [S.l.], v. 21, n. 2, p. 217–244, dez. 2016. Disponível em: <http://bit.ly/2uLpbgL>. Acesso em: 22 jul. 2017.

VICTORINO, M. C. et al. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais concetados. **Informação & Sociedade**, v. 27, n. 1, 2017.

W3C. Resource Description Framework (RDF). 2004. Disponível em: <https://www.w3.org/RDF/>.Acesso em: 28 ago. 2017.

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