ARTIGO

RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação
RDBCI : Digital Journal of Library and Information Science

DOI 10.20396/rdbci.v16i1.8650412

A PRODUÇÃO CIENTÍFICA SOBRE QUALIDADE DE DADOS EM BIG DATA: UM ESTUDO NA BASE DE DADOS WEB OF SCIENCE

THE SCIENTIFIC PRODUCTION ON DATA QUALITY IN BIG DATA: A STUDY IN THE WEB OF SCIENCE DATABASE

LA PRODUCCIÓN CIENTÍFICA SOBRE CALIDAD DE DATOSEN BIG DATA: UNESTUDIO EM LA BASE DE DATOS WEB OF SCIENCE

[1]Priscila Basto Fagundes, [1]Douglas Dyllon Jeronimo de Macedo, [2]Gislaine Parra Freund
[1] Universidade Federal de Santa Catarina
[2] Dígitro Tecnologia

Correspondência

Priscila Basto Fagundes
Universidade Federal de Santa Catarina
Florinópolis, SC
Email: priscila.bfagundes@gmail.com
ORCID:
http://orcid.org/0000-0002-9461-311X

Submetido em: 22/09/2017
Aceito em: 02/11/2017
Publicado em: 09/11/2017

JITA: IE. Data and metadata structure

RESUMO: Cada vez mais o tema big data tem despertado interesse em pesquisadores das mais diferentes áreas do conhecimento, entre eles os cientistas da informação que necessitam compreender seus conceitos e aplicações para poderem contribuir com novas propostas para a gestão das informações geradas a partir dos dados armazenado nestes ambientes.O objetivo deste artigo é apresentar um levantamento das publicações relacionadas a qualidade de dados em big data na base de dados Web of Science até o ano de 2016. Serão apresentados o total de publicações indexadas na base, a quantidade de publicações por ano, o local de publicação das pesquisas e uma síntese dos estudos encontrados. A pesquisa na base de dados foi realizada em julho de 2017 e resultou em um total de 23 publicações. A fim de possibilitar a apresentação de um resumo das publicações neste artigo foram realizadas buscas pelos textos completos de todas as publicações na internet e feita a leitura dos que se encontravam disponíveis. Com este levantamento foi possível concluir que os estudos sobre qualidade de dados em big data tiveram suas publicações a partir de 2013, sendo que a sua maioria apresenta revisões da literatura e poucas propostas efetivas para o monitoramento e gestão da qualidade de dados em ambientes com grandes volumes de dados. Sendo assim, pretende-se com este levantamento contribuir e fomentar novas pesquisas sobre o contexto qualidade de dados em ambientes big data.

PALAVRAS-CHAVE: Qualidade de dados.Big Data.Gestão da Qualidade. Web of Science

ABSTRACT: More and more, the big data theme has attracted interest in researchers from different areas of knowledge, among them information scientists who need to understand their concepts and applications in order to contribute with new proposals for the management of the information generated from the data stored in these environments. The objective of this article is to present a survey of publications about data quality in big data in the Web of Science database until the year 2016. Will be presented the total number of publications indexed in the database, the number of publications per year, the location the origin of the research and a synthesis of the studies found. The survey in the database was conducted in July 2017 and resulted in a total of 23 publications. In order to make it possible to present a summary of the publications in this article, searches were made of the full texts of all the publications on the Internet and read the ones that were available. With this survey it was possible to conclude that the studies on data quality in big data had their publications starting in 2013, most of which present literature reviews and few effective proposals for the monitoring and management of data quality in environments with large volumes of data. Therefore, it is intended with this survey to contribute and foster new research on the context of data quality in big data environments.

KEYWORDS: Data Quality. Big Data. Quality Management. Web of Science

RESUMEN: Cada vez más el tema big data ha despertado interésen investigadores de las más diversas áreas delconocimiento, entre elloslos científicos de lainformación que necesitancomprender sus conceptos y aplicaciones para poder contribuir connuevaspropuestas para lagestión de lasinformacionesgeneradas a partir de losdatosalmacenadosenestosdatoslos ambientes. El objetivo de este artículo es presentar unrelevamiento de laspublicaciones relacionadas conlacalidad de datosen big data enla base de datos Web of Science hasta elaño 2016. Se presentaránel total de publicaciones indexadas enla base, lacantidad de publicaciones al año, el local de origen de lasinvestigaciones y una síntesis de losestudios encontrados. La investigaciónenla base de datos se realizóenjulio de 2017 y resultóenun total de 23 publicaciones. Conelfin de posibilitarlapresentación de unresumen de laspublicaciones, se realizaronbúsquedas por los textos completos de todas laspublicacionesen internet y lalectura de los que se encontrabandisponibles. Con este levantamientofueposible concluir que losestudios sobre calidad de datosen big data tuvieron sus publicaciones a partir de 2013, siendo que sumayoría presenta revisiones de la literatura y pocaspropuestasefectivas para elmonitoreo y gestión de lacalidad de datosen ambientes con grandes volúmenes de datos.

PALABRASCLAVE: Calidad de losdatos. Big Data. Gestión de laCalidad. Web of Science

1 INTRODUÇÃO

Desde o início do século XXI ocorreram mudanças significativas no âmbito das Tecnologias da Informação e Comunicação (TIC), neste contexto pode-se citar a computação em nuvem, internet das coisas e as redes sociais. O acesso e o uso destas tecnologias fizeram com que a quantidade de dados aumentasse de uma forma contínua e a uma velocidade sem precedentes (CAI; ZHU, 2015). De acordo com Furlan e Laurindo (2017), o crescimento dos dados gerados demandou o desenvolvimento de novas soluções e tecnologias que auxiliassem na sua gestão. E diante desta necessidade surge o *big data*, propondo novas abordagens para a geração, seleção e manipulação destes grandes volumes dados. O termo *big data* está relacionado com grandes quantidades de dados, que possuem características distintas, são heterogêneos, providos de diferentes fontes, com controles distribuídos e descentralizados (MCAFEE; BRYNJOLFSSON2012).

Dada a importância das informações disponibilizadas, a qualidade dos dados que geram estas informações tornou-se um dos grandes desafios para as organizações se manterem em um mercado cada vez mais competitivo, sendo quea partir da década de 90 iniciaram-se diversos estudos sobre metodologias e ferramentas para auxiliar no processo de gestão da qualidade dos dados dentro das organizações, e uma dasproposições mais relevantes foi o *Total Data Quality Management* (TDQM), feita por Madnick e Wang em 1992 (ZHU*et al.*, 2012). O programa desenvolvido pelo Massachusetts Instituteof Technology - MIT é baseado na estrutura de Gerenciamento de Qualidade Total (TQM) para melhoria da qualidade no domínio da fabricação, proposto por William Edwards Deming em 1982. Suas pesquisas iniciais desenvolveram um modelo que defende a melhoria contínua da qualidade dos dados, seguindo ciclos de definição, medição, análise e melhoria. A partir do TDQM, várias outras proposições relacionadas as dimensões e atributos da qualidade dos dados foram feitas, porém, a definição de quais critérios a serem adotados depende do contexto em que os mesmos serão aplicados (BATINI*et al.*, 2009).

Procurando estabelecer uma relação dos temas abordados neste artigo com a Ciência da Informação, é válidojulgara importância do envolvimento do profissional de informação nas discussões sobre os temas qualidade de dados e*big data*, uma vez queo uso de dados e informações sempre foi objeto de estudo para a área. Outra questão a se considerar remete à diversidade dos dados disponíveis, uma vez que os mesmossãooriginadosa partir de diferentes fontes, causando uma sobrecarga de informação para a sociedade, gerando inúmeras oportunidades de atuação para os profissionais que atuam na área da gestão da informação (RIBEIRO, 2014).

Desta forma, pretende-se neste artigo contribuir com a geração de novos estudos na área da Ciência da Informação, apresentando um levantamento da produção científica sobre estes assuntos na base de dados Web of Science, fornecendo aos pesquisadores informações que poderão ser utilizadas em futuras pesquisas. Sendo assim, esse artigo tem como objetivo

principal, apresentar um panorama geral das publicações relacionadas à qualidade de dados em ambientes *big data* na base de dados Web of Science, apresentando a quantidade de publicações envolvendo os dois temas, o local de publicação destas pesquisas, uma síntese dos estudos, seus contextos de aplicação eseus autores.

2 QUALIDADE DE DADOS

A partir de uma perspectiva científica, a qualidade dos dados vem sendo abordada em diferentes áreas do conhecimento, como por exemplo, a ciência da informação, a computação, a gestão entre outras. No final da década de 1960, os estatísticos foram os primeiros a investigar os problemas relacionados à qualidade dos dados propondo uma teoria matemática para as duplicidades em conjuntos de dados estatísticos. Estes foram seguidos por pesquisadores na área da gestão, que no início da década de 1980 focavam em como controlar sistemas geradores de dados para detectar e eliminar problemas de qualidade. Somente no início da década de 1990, os cientistas da computação começaram a considerar o problema de definir, medir e melhorar a qualidade dos dados armazenados em bancos de dados e sistemas legados (BATINI; SCANNAPIECA, 2006).

Dentro da ciência da informação, as discussões acerca da qualidade dos dados e informações tiveram início em 1989, em um seminário ocorrido em Copenhage na Dinamarca e promovido pelo NordicConcil for ScientificInformationandResearchLibraries (NORDINFO). A publicação decorrente do encontro tornou-se uma importante referência em relação ao assunto e foi reconhecida como o esforço mais importante de teorização sobre o tema na época (PAIN; NEHMY; GUIMARÃES, 1996).De acordo com Valente e Fujino (2016), embora ainda não haja um consenso sobre o conceito de qualidade, observa-se na ciência da informação diversas proposições de critérios ou atributos para qualificar os dados e as informação. Tais critérios são usados para definir, medir e gerenciar a qualidade e variam de acordo com as abordagens e vertentes sob as quais os estudos são realizados.

Para Wang e Strong (1996), o conceito de qualidade dos dados é considerado multidimensional e as suas dimensões são definidas como um conjunto de atributos de qualidade que representam um único aspecto da qualidade de dados. Observa-se na literatura, tanto estudos que propõem que a qualidade dos dados seja considerada de forma genérica, onde a partir de um determinado modelo a qualidade pode ser gerida independente do domínio de aplicação, como também pesquisas que procuraram definir e operacionalizar as dimensões de qualidade de dados específicas para determinadoscontextos. Essa diversidade poderá ser observada na sessão 5 deste artigo, onde serão apresentados e resultados obtidos no levantamento realizado.

Wang e Strong (1996), conduziram o que foi considerada a primeira pesquisa empírica com o propósito de identificar as dimensões da qualidade dos dados. Os autores propuseram incialmente um quadro conceitual que inclui os seguintes aspectos:

- o usuário deve ser capaz de obter os dados, o que significa que os dados devem ser acessíveis;
- é fundamental que o usuário consiga compreender a sintaxe e a semântica dos dados;
- os dados devem ser úteis;
- os dados devem possuir credibilidade para o usuário.

Com base nesses aspectos, o estudo foi dividido em duas etapas, sendo que a primeira teve como propósito, gerar uma lista de dimensões da qualidade dos dados a partir da percepção dos usuários – com 137 participantes, e a segunda teve o objetivo de identificar a importância de cada dimensão levantada na primeira etapa, essa com aproximadamente 1.480 participantes. O resultado preliminar, apontou para um total de 20 dimensões, consideradas as mais relevantes na visão dos usuários, porém essa quantidade ainda era considerada alta para fins de avaliação da qualidade. Foi então, realizada uma nova análise com o objetivo de classificar essas dimensões em categorias, definidas como: Intrínseca (precisão dos dados); Contextual (relevância dos dados); Representacional (forma de representação dos dados) e; Acessibilidade (formas de acesso aos dados) (WANG; STRONG, 1996). Como resultado final, chegou-se a um conjunto de 15 dimensões:

Quadro 1. Dimensões da qualidade dos dados

CATEGORIAS	ATRIBUTOS
Intrínseca	Preciso, objetivo, com credibilidade, fidedigno
Contextual	Relevante, com valor agregado, atualizado, completo, com valor apropriado
Representacional	Interpretável, com facilidade de entendimento, representação concisa, com representação consistente
Acessibilidade	Acessível e seguro

Fonte: (WANG e STRONG, 1996)

Vale ressaltar, que a partir da pesquisa de Wang e Strong (1996) várias outras proposições relacionadas as dimensões e atributos da qualidade dos dados foram feitas. No entanto, Batini*et al.* (2009) argumenta que não existe um acordo quanto ao conjunto de dimensões que definem a qualidade, nem dos dados nem do significado exato de cada dimensão.

3 BIG DATA

Para Erl,Khattak e Buhler (2016), *big data* é um campo que se dedica à análise, ao processamento e armazenamento de grandes *datasets*[1], e que as suas soluções e práticas são geralmente necessárias quando as tecnologias e técnicas tradicionais não são suficientes para a execução destas atividades. Para os autores, *big data* não é apenas uma tecnologia, é também sobre como as tecnologias podem impulsionar uma organização

Cada vez mais as organizações estão gerando enormes quantidades de dados que são provenientes de fontes distintas e armazenados de diferentes maneiras, o que demanda um processo de gestão específico para garantir a sua qualidade. Vianna, Dutra e Frazzon (2016, p.193), enfatizam a importância de se fazer uma gestão efetiva e a [...] "necessidade de transformar esses dados em informações de qualidade, que possam ser utilizadas para direcionar os negócios e as estratégias das organizações, minimizando riscos, e apoiando o processo de tomada de decisões".

Existem na literatura, diferentes pontos de vista em relação as características que compõem um ambiente *big data*. Três delas foram inicialmente identificadas por Doug Laney no início de 2001, quando o autor publicou um artigo descrevendo o impacto do volume, da velocidade e da variedade dos dados em *data warehouses* corporativos (LANEY, 2001). Anos mais tarde novas características foram incorporadas ao conjunto de aspectos relacionados a este conceito, são elas: veracidade, valor, variabilidade (ZIKOPOULOS *et al*., 2012; GANDOMI; HAIDER, 2015).Buscando auxiliar em um melhor entendimento acerca das características dos ambientes *big data*, a seguir será apresentadoo conceito de cada uma delas.

- Volume: refere-se a grandes quantidades de dados e informações que são geradas a partir de fontes variadas. De acordo com Erl,Khattak e Buhler (2016), o volume de dados que é processado pelas soluções *big data*é substancial e cada vez maior e impõem exigências distintas de armazenamento e processamento, apresentando um grande desafio às estruturas de TI convencionais, pois grandes volumes de dados requerem armazenamento escalonável e uma abordagem distribuída para possibilitar a consulta aos mesmos.

- Variedade: diz respeito a diversidade dos dados e informações. Kaisler*et al.* (2013), afirma que a partir de uma perspectiva analítica, fazer a gestão da variedade de dados é provavelmente o maior obstáculo para utilização efetiva de grandes volumes de dados. Formatos de dados incompatíveis, estruturas de dados não alinhadas e semânticas inconsistentes representam desafios significativos que se não vencidos podem levar ao insucesso de um projeto.

[1]*Datasets* são coleções ou agrupamentos de dados relacionados, onde cada grupo ou membro do grupo compartilha as mesmas propriedades ou atributos (ERL;KHATTAK; BUHLER, 2016).

- Velocidade: possui relação com o tempo de resposta para determinada requisição. Com as comunicações em tempo real, cada vem mais vem sendo possível atingir uma maior velocidade para troca de dados e informação (RIBEIRO, 2014). Mcafee e Brynjolfsson (2012) afirmam que para muitas aplicações, a velocidade de criação de dados é ainda mais importante do que o volume dos mesmos, uma vez que as informações em tempo real ou quase em tempo real, tornam possível que uma empresa seja mais ágil que seus concorrentes.

- Veracidade: está relacionada à qualidade ou à confiabilidade dos dados. Uma definição acerca desta característica foi apresentada por Erl,Khattak e Buhler (2016), que afirmam que os dados precisam ser avaliados quanto à confiabilidade, o que pode demandar atividades específicas para identificar os que não atendem a essa premissa e removê-los dos *datasets*. Um dado considerado de má qualidade é aquele que não pode ser convertido em informação e, portanto, não tem valor, enquanto que dados de qualidade possuem valor e geram informações significativas.

- Valor: está relacionado com o retorno do investimento e é o resultado da combinação dos aspectos citados anteriormente. Esta característica está intuitivamente relacionada com a característica de veracidade, pois quanto maior a qualidade dos dados, mais valor ela possui para o negócio (KAISLER *et al.*, 2013). Da mesma forma, pode-se considerar que valor e velocidade possuem relações inversas, pois quanto mais tempo se leva para que os dados sejam transformados em informações relevantes, menos valor terá para o negócio, uma vez que resultados obsoletos prejudicam a qualidade e a rapidez na tomada de decisão.

- Variabilidade. refere-se aos dados que estão em constante variação, como por exemplo, dados meteorológicos. Gandomi e Haider (2015), afirmam que a variabilidade possui relação com a variação nos fluxos de dados.

Com as características apresentadas é possível determinar a noção do termo *big data*. Porém, especialistas acreditam que não há a necessidade da existência de todos os fatores ao mesmo tempo para um ambiente seja considerado *big data*, visto que existem casos em que há um maior destaque para uma ou outra característica, e em outros casos elas não são identificadas em sua totalidade

4 PROCEDIMENTOS METODOLÓGICOS

A consulta à Web of Science foi realizada em julho de 2017 e utilizou-se a string: *("data quality" and "big data")*no o campo de pesquisa "Título". Quanto ao tipo de documento foram selecionados apenas os artigos e os *proceedingspaper* indexados pela base, excluindo os documentos caracterizados por *book review*, *editorial material* e *review* entre

outros de menor ocorrência até o ano de 2016, resultando em 23 artigos sobre os termos pesquisados.Após a identificação das publicações foram realizadas buscas na internet, através do acesso institucional da Universidade Federal de Santa Catarina (UFSC), pelos textos completos dos artigos e dos 23 estudos encontrados, 6 deles não estavam disponíveis de forma gratuita pelo acesso institucional. Por este motivo os seus resumos não serão apresentados neste artigo, porém os mesmos farão parte da análise quantitativa apresentada na próxima seção.

Vale ressaltar que em um primeiro momento, foi definido que a busca contemplaria o campo de pesquisa Tópico, que engloba a consulta ao título, ao resumo e as palavras-chaves, o que resultou em um total de 158 publicações, para as mesmas condições de busca. Porém, após uma análise preliminar nos títulos das publicações, constatou-se que a grande maioria não focava especificamente em qualidade de dados em ambientes *big data*, sendo assim, com o objetivo de contemplar somente os estudos relacionados a estas duas áreas, optou-se por utilizar apenas o campo Título.

Outra questão a ser levada em consideração, foi a opção de utilizar apenas o termo relacionado a qualidade de dados e não fazer uso, também, do termo qualidade da informação. Alguns autores não apresentam distinção entre os conceitos de qualidade de dados e qualidade da informação, tratando ambos como tendo o mesmo significado. Porém, acredita-se que os dados são um conjunto de valores ou ocorrências que após processados geram as informações (SOMASUNDARAM; SHRIVASTAVA, 2011), sendo assim no presente estudo, será acatado que dado e informação possuem significados diferentes e para este levantamento serão consideradas apenas as publicações que fazem referência à qualidade de dados.

A base de dados Web of Science foi escolhida por ser considerada uma base interdisciplinar e um importante indexador de periódicos científicos, possibilitando que sejam feitas análises quantitativas nos resultados obtidos através das consultas (PORTAL DE PERIÓDICOS DA CAPES/MEC, 2014).

5 RESULTADOS

A pesquisa realizada na Web of Science no dia 05/07/2017 para os termos *"data quality" and "big data"* recuperou no total 17 *proceedingspapers* e 6 artigos, totalizando 23 publicações. Conforme apresentado no Gráfico 1, as publicações sobre o tema iniciaram em 2013 com 3 publicações neste ano, sendo que o maior número de estudos publicados, 9 no total, foi em 2015.

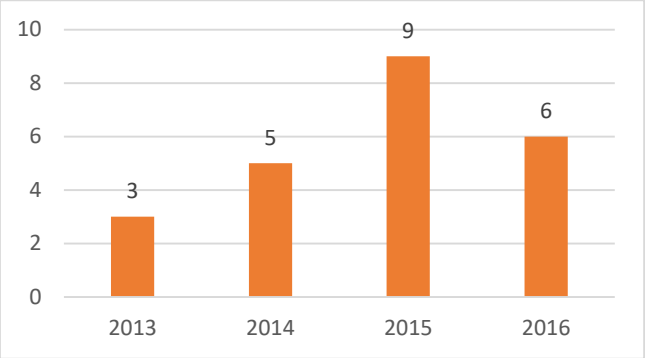

Gráfico 1: Número de publicações sobre qualidade de dados em big data

Fonte: WoS 2017

A maioria dos estudos teve origem nos EUA, com 8 das 23 publicações, Espanha e Itália contabilizaram duas cada um e países como Brasil, Alemanha, Emirados Árabes, Suíça, França, Canadá, Holanda, Coreia do Sul, Escócia, Arábia Saudita, China, Austrália e Ilhas Maurício obtiveram uma publicação sobre o tema.

O Quadro 1, apresenta o título, o(s) autor(es), o ano de todas das publicações encontradas e os contextos das pesquisas. Dos 23 estudos, 6 não possuem seus textos completos disponíveis pelo acesso institucional da Universidade Federal de Santa Catarina (UFSC) são eles: 3, 4, 7, 9, 12 e 16, e por este motivo não terão suas sínteses apresentadas.

Quadro 1: Publicações sobre qualidade de dados e big data.

	TÍTULO	CONTEXTO
1	Pay-as-you-go data quality improvement for medical centers (ENDLER; BAUMGAERTEL; LENZ, 2013)	Gestão de dados financeiros na área da saúde
2	Information governance, big data and data quality (FREITAS *et al*., 2013)	Governança da informação
3	Big data quality analysis (QIN, 2013)	-
4	A data quality in use model for big data (CABALLERO, 2014)	-
5	Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications (HAZEN *et al*., 2014)	Gestão de dados na cadeia de suprimentos
6	Data quality management, data usage experience and acquisition intention of big data analytics (KWON; LEE; SHIN, 2014)	Gestão da qualidade de dados
7	Data quality issues concerning statistical data gathering supported by big data technology (MASLANKOWSKI, 2014)	-
8	Data quality: the other face of big data (SAHA; SRIVASTAVA, 2014)	Gestão da qualidade de dados
9	Framework for social media big data quality analysis (AL-HAJJAR, 2015)	-
10	From data quality to big data quality (BATINI *et al*.,	Gestão da qualidade de dados

	2015)	
11	Big data, big data quality problem (BECKER; MCMULLEN; KING, 2015)	Gestão da qualidade de dados
12	Scalable data quality for big data: the pythia framework for handling missing values (CAHSAI *et al*., 2015)	-
13	Overview of data quality challenges in the context of big data (JUDDOO, 2015)	Gestão da qualidade de dados
14	Taking a 'big data' approach to data quality in a citizen science project (KELLING *et al*., 2015)	Gestão da informação no monitoramento de aves
15	Data quality issues in big data (RAO; GUDIVADA; RAGHAVAN, 2015)	Gestão daqualidade de dados biológicos
16	I8k vertical bar dq-bigdata: i8k architecture extension for data quality in big data (RIVAS, 2015)	-
17	Computing data quality indicators on big data streams using a cep (YANG; SILVA; PICARD, 2015)	Gestão de dados em cidades inteligentes
18	Big data quality: a roadmap for open data (CIANCARINI, 2016)	Gestão da qualidade de dados abertos
19	Data quality: experiences and lessons from operationalizing big data (GANAPATHI, 2016)	Gestão da qualidade de dados
20	Antecedents of big data quality an empirical examination in financial service organizations (HARYADI *et al*., 2016)	Gestão da qualidade de dado financeiros
21	A data quality in use model for big data (MERINO *et al*., 2016)	Gestão da qualidade de dados
22	Big data quality - whose problem is it? (SADIQ; PAPOTTI, 2016)	Gestão da qualidade de dados
23	Big data quality: a quality dimensions evaluation (TALEB *et al*., 2016)	Gestão da qualidade de dados

Fonte: Web ofScience, 2017

A seguir serão apresentados os objetivos e uma breve descrição das publicações encontradas.

O estudo de Hazen*et al*. (2014), publicado no *InternationalJournalofProductionEconomics* em 2014 objetiva a proposição de um modelo na perspectiva da ciência dos dados, análise preditiva e big data para a melhoria contínua na produção de dados, com foco na gestão de cadeias de suprimentos. O modelo proposto é aplicado em um estudo de caso, bem como apresentados os seus resultados.

A partir de uma pesquisa empírica, Kwon, Lee e Shin(2014) apresenta um mapeamento das questões que levam as organizações a adotarem cenários *big data*. No estudo, é detalhada a metodologia utilizada para a aplicação da pesquisa e descritos os resultados obtidos, mostrando que principalmente fatores externos como a situação econômica global, a pressão do mercado para oferta de melhores serviços ou produtos e novas oportunidades de negócios, são fatores determinantes para adoção de novas estratégias como *big data* e *big data analytics*. Este artigo foi publicado no

InternationalJournalofInformation Management . Outro estudo com domínio de aplicação bastante específico foi publicado por Kelling*et al*. (2015) no AMBIO. O artigo apresenta uma proposta para a utilização de *big dada* em um projeto de monitoramento de aves denominado eBird, e tem como objetivo aumentar a qualidade dos dados gerados e disponibilizados aos pesquisadores do projeto.

O artigo publicado na *Future Generation Computer Systems - The InternationalJournalofEscience*por Merino *et al*. (2016), considera a qualidade de dados em *big data* através da perspectiva do usuário e propõe um modelo denominado "'*3As Data Quality-in-Usemodel''*. O mesmo é baseado nas normas ISO/IEC 25012 e ISO/IEC 25024, e os pesquisadores sugerem que o conjunto de dimensões que compõem o modelo sejam divididas em 3 grupos: adequação contextual, adequação, temporal e adequação operacional.

Algumas publicações são voltadas para a identificação de problemas relacionados a qualidade de dados em *big data*. Batini*et al*. (2015), apresentam em seu artigo publicado no *JournalofDatabase Management*, uma análise sobre os problemas relacionados a qualidade de dados e propõem um *framework* conceitual para monitorar a qualidade de acordo com três pontos, considerados poreles, relevantes em ambientes *big data*, são eles: o tipo do dado, a sua origem e o domínio da aplicação.

Becker, Mcmullen e King(2105) e Ganapathi e Chen(2016), apresentaram estudos envolvendoum levantamento dos problemas relacionados a qualidade de dados propondo um conjunto de soluções para os mesmos. O artigo de Rao, Gudivada e Raghavan(2015), contempla as dificuldades identificadas durante a integração e o compartilhamento dos dados, tendo como foco dados biológicos. Um estudo semelhante ao anterior, é apresentado em Haryadi*et al.* (2016), porém nesta publicação o objetivo foi identificar os problemas relacionados a qualidade de dados em *big data* no contexto das instituições financeiras, onde foram definidas 11 dimensões da qualidade de dados a serem consideradas no domínio em questão e realizada uma avaliação em 3 instituições, descrevendo a metodologia aplicada, bem como os resultados obtidos. Os 4 últimos trabalhos foram apresentados na IEEE *InternationalConferenceon Big Data* nos anos de 2015 e 2016.

Também na área da saúde e objetivando a melhoria contínua, Endler,Baumgaertel e Lenz(2013) apresentam na *ConferenceoneHealth - Health InformaticsMeetseHealth*de 2013, um modelo baseado do TDQM a ser adotado no monitoramento da qualidade dos dados financeiros de centros de saúde, sendo que, no momento em que a proposta foi publicada, o modelo ainda se encontrava em fase de validação, não apresentado resultados efetivos.

No contexto das "s*martcities*", Yang, Silva ePicard(2015) propõem um *framework*que foi apresentado no *International Workshop onComputationalIntelligence for MultimediaUnderstanding* em 2015, para calcular indicadores de qualidade genéricos nos fluxos de dados dos medidores inteligentes, com base na tecnologia de processamento de

eventos complexos *ComplexEventProcessing (CEP),* na época da publicação o framework proposto já se encontrava em utilização na França.

A qualidade dos dados abertos *(Open Data)* é discutida no estudo de Ciancarini, Poggi e Russo (2016) e apresentada no *2nd IEEE InternationalConferenceon Big Data Computing Service andApplications*. Segundo os autores, garantir a qualidade dos dados abertos é um dos grandes desafios deste movimento e o objetivo principal do estudo é realizar uma análise dos dados disponibilizados por instituições ligadas ao governo italiano e o modelo utilizado na avaliação é o ISO/IEC 25012.

O estudo de Taleb*et al*. (2016), sugere que a avaliação da qualidade dados em *big data* seja feita por amostragem. Os pesquisadores desenvolveram um algoritmo para avaliação da qualidade que é aplicado em um estudo de caso, e argumentam que a utilização dessa abordagem é eficiente, uma vez que reduz o tempo e os recursos de computação envolvidos. O estudofoiapresentado no *Int IEEE Conferences nn Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress em 2016.*

Juddoo (2015), apresenta na *InternationalConferenceonComputing, Communication and Security*, uma revisão da literatura sobre gestão da qualidade de dados em diferentes contextos e Freitas *et al.* (2013), apresentam uma breve discussão a respeito de "*informationgovernance*". A publicação traz conceitos sobre *big data* e qualidade de dados de uma maneira sucinta e breve, o estudo foi apresentado no IEEE *16th InternationalConferenceonComputational Science andEngineering.*

E por fim, as publicações de Saha eSrivastava(2014) e SadiqePapotti(2016), não tratam especificamente de um estudo e sim da programação de uma palestra sobre os desafios da gestão da qualidade de dados em *big data* e de uma mesa redonda sobre as responsabilidades acerca da qualidade em *big data* respetivamente, ambas apresentadas na *InternationalConferenceon Data Engineering* em 2014 e 2016.

4 CONSIDERAÇÕES FINAIS

Conforme apresentado neste artigo, pode-se concluir que as discussões relacionadas ao tema qualidade de dados em ambientes *big data*são de certa forma recentes, uma vez que a primeira publicação na base de dados Web of Science foi no ano de 2013. Outra questão relevante são os contextos onde ocorreram as pesquisas, percebe-se que a maioria das publicações apresentam estudos relacionados a gestão da qualidade de dados em ambientes *big data* de forma genérica, sem especificidades no domínio de aplicação, pois foram poucas as pesquisas direcionadas a um determinado contexto, como por exemplo o financeiro. Em relação ao local das publicações, é observado que os periódicos ou eventos onde ocorreram as

mesmas apresentam uma tendência a estarem relacionados à área da computação, o que é compreensível pelo fato do tema *big data*possuir uma forte relação com esta área.

A partir das análisesnos resultados obtidos, é possível afirmar a existência de oportunidades de pesquisas sobre o tema qualidade de dados em *big data* dentro da Ciência da Informação, que poderá contribuir com a aplicação de conceitos sobre dados e informação que vão além da tecnologia, como por exemplo, gestão e fluxo dos mesmos. Por entender que os dados possuem características e necessidades inerentes ao contexto a que pertencem, outras sugestões de trabalhos futuros, seriam pesquisas que contemplem modelos a serem aplicados em áreas específicas do conhecimento, bem como a realização de levantamentos como este em outras bases de dados, que venham a somar com as publicações apresentadas neste artigo.

REFERÊNCIAS

BATINI, Carlo; SCANNAPIECA, Monica. **Data Quality: Concepts, Methodologies and Techniques**. New York. Springer, 2006

BATINI, Carlo *et al*. Methodologies for Data Quality Assessment and Improvement. **ACM ComputingSurveys**, n.3, v.41, 2009, p. 1-52. Disponível em: <http://dl.acm.org/citation.cfm?id=1541883>. Acessoem: 25 mai. 2017.

BATINI, Carlo. *et al.* From Data Quality to Big Data Quality. **Journal of Database Management**, v. 26, n. 1, 2015, p. 60–82. Disponível em: <https://www.researchgate.net/publication/283681085_From_Data_Quality_to_Big_Data_Quality>. Acessoem: 7 jul. 2017.

BECKER, David; MCMULLEN, Bill; KING, Trish Dunn. Big Data, Big Data Quality Problem. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2015, Santa Clara. **Anais eletrônicos**... Santa Clara: 2015. p.2644-2653 Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364064>. Acessoem: 7 jul. 2017.

CAI, Li; ZHU, Yangyong. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. **Data Science Journal**, v. 14, n. 0, 2015, p. 2. Disponível em: <http://datascience.codata.org/article/10.5334/dsj-2015-002/>. Acessoem: 15 jun. 2017.

CIANCARINI, Paolo; POGGI, Francesco; RUSSO, Daniel. Big Data Quality: a Roadmap for Open Data. 2ND IEEE INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING SERVICE AND APPLICATIONS, 2., 2016, Oxford. **Anais eletrônicos**… Praga: 2016. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7474375>. Acessoem: 7 jul. 2017.

ENDLER, Gregor; BAUMGAERTEL, Philipp; LENZ, Richard. Pay-as-you-go data quality improvement for medical centers. In: CONFERENCE ON EHEALTH - HEALTH INFORMATICS MEETS EHEALTH, 2013, Vienna. **Anais eletrônicos**… Vienna: 2013.

p.13-18. Disponível em: <http://www.ehealth20xx.at/wp-content/uploads/scientific-papers/2013/endler.pdf>. Acessoem: 7 jul. 2017.

ERL, Thomas; KHATTAK, Wajid; BUHLER, Paul. **Big Data Fundamentals: Concepts, Drivers & Techniques**. Boston: Prentice Hall, 2016.

FREITAS, Patrícia Alves de *et al.* Information Governance, Big Data and Data Quality. In: IEEE 16TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ENGINEERING (CSE), 16., 2013, Sydney. **Anais eletrônicos**… Sydney: 2013. p.1142-1143. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6755349>. Acesso em 07 jul. 2017.

FURLAN, PatriciaKuzmenko; LAURINDO, Fernando José Barbin. Agrupamentos epistemológicos de artigos publicados sobre big data analytics. **Transinformação**, v. 29, n. 1, 2017, p. 91-100. Disponível em: <http://www.scielo.br/pdf/tinf/v29n1/0103-3786-tinf-29-01-00091.pdf>. Acessoem: 21 abr. 2017.

GANAPATHI, Archana; CHEN, Yanpei. Data Quality: Experiences and Lessons from Operationalizing Big Data. 4TH IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 4., 2016, Washington. **Anais eletrônicos**… Washington: 2016. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7840769>. Acessoem: 7 jul. 2017.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management**, v. 35, n. 2, 2015, p. 137–144. Disponívelem: <http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>. Acessoem: 21 abr. 2017.

HARYADI, AdiskaFardani*et al.* Antecedents of Big Data Quality An Empirical Examination in Financial Service Organizations. 4TH IEEE INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 4., 2016, Washington. **Anais eletrônicos**… Washington: 2016. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7840595>. Acessoem: 7 jul. 2017.

HAZEN, Benjamin T. *et al.* Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. **International Journal of Production Economics**, v. 154, 2014, p. 72–80. Disponívelem: <http://www-sciencedirect-com.ez46.periodicos.capes.gov.br/science/article/pii/S0925527314001339?via%3Dihub>. Acessoem: 7 jul. 2017.

JUDDOO, Suraj. Overview of data quality challenges in the context of Big Data. In: INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND SECURITY (ICCCS), 2015, Pamplemousses. **Anais eletrônicos**… Pamplemousses : 2015. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7374131>. Acessoem: 7 jul. 2017.

KAISLER, Stephen *et al.* Big Data: Issues and Challenges Moving Forward. In: XLVI HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 46., Maui, 2013.

Anais eletrônicos... Maui, 2013. p.995-1004. Disponível em: <https://www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892a995.pdf>. Acessoem: 22 abr. 2017.

KELLING, Steve *et al.* Taking a `Big Data' approach to data quality in a citizen science project. **AMBIO**, v. 44, n. 4, 2015, p. S601–S611. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623867/>. Acessoem: 7 jul. 2017.

KWON, Ohbyung; LEE, Namyeon; SHIN, Bongsik. Data quality management, data usage experience and acquisition intention of big data analytics. **International Journal of Information Management**, v. 34, n. 3, 2014, p. 387–394. Disponívelem: <http://www-sciencedirect-com.ez46.periodicos.capes.gov.br/science/article/pii/S0268401214000127?via%3Dihub>. Acessoem: 7 jul. 2017.

LANEY, Doug. Application Delivery Strategies. **META Group**, 2001. Disponível em: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acessoem: 7 jul. 2017.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data. The management revolution. **Harvard BuinessReview**, v. 90, n. 10, 2012 p. 61–68. Disponível em: <https://hbr.org/2012/10/big-data-the-management-revolution>. Acesso em: 22 abr. 2017.

MERINO, Jorge *et al.*A Data Quality in Use model for Big Data. **Future Generation Computer Systems**, v. 63, 2016, p.123-130. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X15003817/>. Acesso em: 07 jul. 2017.

PAIM, Isis; NEHMY, Rosa Maria Quadros, GUIMARÃES, César Geraldo. Problematização do conceito "Qualidade" da Informação. **Perspectivas em Ciência da Informação**, v. 1, n. 1, 1996, p. 111–119. Disponível em <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/8/27>. Acesso em: 30 mar. 2017.

PORTAL DE PERIÓDICOS DA CAPES/MEC. Disponível em:<http://www.periodicos.capes.gov.br/?option=com_pcollection&mn=70&smn=79&cid=81>. Acessoem: 07 jun. 2017.

RAO, Dhana; GUDIVADA, Venkat N.; RAGHAVAN, Vijay V. Data Quality Issues in Big Data. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2015, Santa Clara. **Anais eletrônicos**... Santa Clara: 2015. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364065>. Acesso em: 7 jul. 2017.

RIBEIRO, Claudio José Silva. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia**, v. 1, n. 1, 2014, p. 96–105. Disponível em: <http://periodicos.ufpb.br/index.php/itec/article/view/19380/11156>. Acessoem: 19 abr. 2017.

SADIQ, Shazia; PAPOTTI, Paolo. Big Data Quality - Whose problem is it? 32ND IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE), 32., 2016, Helsinki. **Anais eletrônicos**… Helsinki: 2016. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7498367>. Acessoem: 07 jul. 2017.

SAHA, Barna; SRIVASTAVA, Divesh. Data Quality: The other Face of Big Data. In: IEEE 30TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE), 30., 2014, Chicago. **Anais eletrônicos**…Chicago: 2014. Disponível em: <https://people.cs.umass.edu/~barna/paper/ICDE-Tutorial-DQ.pdf>. Acesso em: 7 jul. 2017.

SOMASUNDARAM, G.; SHRIVASTAVA, Alok.**Armazenamento e Gerenciamento de Informações**: Como armazenar, gerenciar e proteger informações digitais. Porto Alegre: Bookman. 2011. 472p.

TALEB, Ikbal*et al.* Big Data Quality: A Quality Dimensions Evaluation. 13TH IEEE INT CONF ON UBIQUITOUS INTELLIGENCE AND COMP, 13., 2016, Toulouse. **Anais eletrônicos**… Toulouse: 2016. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7816918>. Acesso em: 7 jul. 2017.

VALENTE, Nelma T. Zubek; FUJINO, Asa. Atributos e dimensões de qualidade da informação nas Ciências Contábeis e na Ciência da Informação: um estudo comparativo**.** **Perspectivas em Ciência da Informação**, v. 21, n. 2, 2016, p. 141–167. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2530/1761>. Acesso em: 16 mar. 2017.

VIANNA, William Barbosa; DUTRA, Moisés Lima; FRAZZON, Enzo Morosini. Big data e gestão da informação: modelagem do contexto decisional apoiado pela sistemografia. **Informação & Informação**, v. 21, n. 1, 2016, p. 185. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/23327/18993>. Acessoem: 21 abr. 2017.

WANG, Richard Y.; STRONG, Diane M. Beyond Accuracy: What Data Quality Means to Data Consumers. **Journal of Management Information System**, v.12, n.4, 1996, p.5-34. Disponívelem: <http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf>. Acesso em: 16 abr. 2017.

YANG, Wenlu; SILVA, Alzennyr Da; PICARD, Marie-Luce. Computing Data Quality Indicators On Big Data Streams Using A Cep. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL INTELLIGENCE FOR MULTIMEDIA UNDERSTANDING (IWCIM), 2015, Praga. **Anais eletrônicos**… Praga: 2015. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7347061>. Acessoem: 10 jul. 2017

ZIKOPOULOS, Paul. *et al.***Understanding Big Data**: Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw-Hill, 2012.

ZHU, Hongwei*et al.* **Data and Information Quality Research:** Its Evolution and Future. MIT: Cambridge, 2012. Disponível em:<http://web.mit.edu/smadnick/www/wp/2012-13.pdf>. Acesso em: 10 jul. 2017.