ARTICLE

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

DOI 10.20396/rdbci.v16i1.8650412

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

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

LA PRODUCCIÓN CIENTÍFICA SOBRE CALIDAD DE DATOSEN BIG DATA: UNESTUDIOENLA 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 - Brazil.
Email: priscila.bfagundes@gmail.com
ORCID:
http://orcid.org/0000-0002-9461-311X

Submitted: 22/09/2017
Accepted: 02/11/2017
Published:12/12/2017

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 INTRODUCTION

Since the beginning of the 21st century, there have been significant changes in the scope of Information and Communication Technologies (ICT), in this context we can mention cloud computing, internet of things and social networks. Access to and use of these technologies has made the amount of data increase continuously and at an unprecedented rate (CAI, ZHU, 2015). According to Furlan and Laurindo (2017), the growth of the data generated demanded the development of new solutions and technologies that would aid in its management. And in the face of this need arises the big data, proposing new approaches for the generation, selection and manipulation of these large volumes. The term big data is related to large amounts of data, which have distinct characteristics, are heterogeneous, provided from different sources, with distributed and decentralized controls (MCAFEE; BRYNJOLFSSON 2012).

Given the importance of the information available, the quality of the data that generates this information has become one of the great challenges for the organizations to remain in an increasingly competitive market, and since the 90's several studies on methodologies and tools to assist in the process of data quality management within organizations, and one of the most relevant propositions was Total Data Quality Management (TDQM), made by Madnick and Wang in 1992 (ZHU et al., 2012). The program developed by the Massachusetts Institute of Technology (MIT) is based on the Total Quality Management (TQM) framework for quality improvement in manufacturing, proposed by William Edwards Deming in 1982. His initial research developed a model that advocates continuous improvement of data quality, following cycles of definition, measurement, analysis and improvement. From the TDQM, several other propositions related to the dimensions and attributes of the data quality were made, however, the definition of which criteria to adopt depends on the context in which they will be applied (BATINI *et al.*, 2009).

In order to establish a relationship between the topics covered in this article and the Information Science, it is valid to judge the importance of the involvement of the information professional in the discussions on the issues of data quality and big data, since the use of data and information has always been an object to the area. Another issue to consider is the diversity of available data, since they originate from different sources, causing an overload of information for society, generating numerous opportunities for professionals working in the area of information management (RIBEIRO, 2014).

In this way, this article intends to contribute with the generation of new studies in the area of Information Science, presenting a survey of the scientific production on these subjects in the Web of Science database, providing to the researchers information that can be used in future researches. Thus, this article has as main objective, to present an overview of publications related to data quality in big data environments in the Web of Science database,

presenting the number of publications involving both themes, the place of publication of these researches, a synthesis of the studies, their contexts of application and their authors.

2 DATA QUALITY

From a scientific perspective, data quality has been approached in different areas of knowledge, such as information science, computing, management and others. In the late 1960s, statisticians were the first to investigate problems related to data quality by proposing a mathematical theory for duplicities in statistical data sets. These were followed by researchers in management who in the early 1980s focused on how to control data-generating systems to detect and eliminate quality problems. Only in the early 1990s did computer scientists begin to consider the problem of defining, measuring and improving the quality of data stored in databases and legacy systems (Bannini and Scannapieeca, 2006).

Within the field of information science, discussions about data quality and information began in 1989 at a seminar held in Copenhagen, Denmark, and promoted by the Nordic Concil for Scientific Information and Research Libraries (NORDINFO). The publication resulting from the meeting became an important reference on the subject and was recognized as the most important theorizing effort on the subject at the time (PAIN, NEHMY, GUIMARÃES, 1996). According to Valente and Fujino (2016), although there is still no consensus on the concept of quality, it is observed in the science of information several propositions of criteria or attributes to qualify data and information. These criteria are used to define, measure and manage quality and vary according to the approaches and strands under which the studies are carried out.

For Wang and Strong (1996), the concept of data quality is considered multidimensional and its dimensions are defined as a set of quality attributes that represent a single aspect of data quality. It is observed in the literature, both studies that propose that data quality be considered in a generic way, where from a given model the quality can be managed independently of the application domain, as well as research that seeks to define and operationalize the dimensions of specific data for certain contexts. This diversity can be observed in section 5 of this article, where they will be presented and results obtained in the survey carried out.

Wang and Strong (1996) conducted what was considered the first empirical research in order to identify the dimensions of data quality. The authors first proposed a conceptual framework that includes the following aspects:

- the user must be able to obtain the data, which means that the data must be accessible;
- it is fundamental that the user is able to understand the syntax and semantics of the data;

• the data should be useful;
• the data must have credibility to the user.

Based on these aspects, the study was divided into two stages, the first of which was to generate a list of data quality dimensions based on the perception of the users - with 137 participants, and the second one was to identify the importance of each dimension raised in the first stage, with approximately 1,480 participants. The preliminary result, pointed to a total of 20 dimensions, considered the most relevant in the view of the users, but this quantity was still considered high for quality evaluation purposes. A new analysis was carried out with the objective of classifying these dimensions into categories, defined as: Intrinsic (data precision); Contextual (relevance of data); Representational (form of representation of data) and; Accessibility (forms of access to data) (WANG; STRONG, 1996). As a final result, a set of 15 dimensions was achieved:

Chart 1. Dimensions of data quality

CATEGORIES	ATRIBUTIONS
Intrinsic	Accurate, objective, credible, reliable
Contextual	Relevant, with added value, updated, complete, with appropriate value
Representational	Interpretable, with ease of understanding, concise representation, with consistent representation
Acessibility	Accessible and safe

Source: (WANG and STRONG, 1996)

It is noteworthy that from the research of Wang and Strong (1996) several other propositions related to the dimensions and attributes of data quality were made. However, Batini et al. (2009) argues that there is no agreement on the set of dimensions that define the quality, neither the data nor the exact meaning of each dimension.

3 BIG DATA

For Erl, Khattak, and Buhler (2016), big data is a field that deals with the analysis, processing and storage of large datasets[1], and that their solutions and practices are often necessary when traditional technologies and techniques are not sufficient for execution of these activities. For the authors, big data is not just a technology, it is also about how technologies can boost an organization

[1] *Datasets* are collections or related data groups, where each group or group member shares the same properties or attributes (ERL; KHATTAK; BUHLER, 2016).

Increasingly organizations are generating huge amounts of data that come from different sources and stored in different ways, which demands a specific management process to ensure its quality. Vianna, Dutra, and Frazzon (2016, p.193), emphasize the importance of effective management and the "need to transform this data into quality information that can be used to direct business and strategies of organizations, minimizing risks, and supporting the decision-making process. "

There are in the literature, different points of view regarding the characteristics that make up a big data environment. Three of them were initially identified by Doug Laney in early 2001, when the author published an article describing the impact of volume, speed and variety of data on enterprise data warehouses (LANEY, 2001). Years later new features were incorporated into the set of aspects related to this concept, they are: veracity, value, variability (ZIKOPOULOS et al., 2012; GANDOMI; HAIDER, 2015). Seeking to aid in a better understanding of the characteristics of big data environments, the following will present the concept of each of them.

• Volume: refers to large amounts of data and information that are generated from a variety of sources. According to Erl, Khattak and Buhler (2016), the volume of data that is processed by big data solutions is substantial and increasing and imposes different storage and processing requirements, presenting a great challenge to conventional IT structures, since large volumes require scalable storage and a distributed approach to enable them to be consulted.

• Variety: refers to the diversity of data and information. Kaisler et al. (2013), states that from an analytical perspective, managing data variety is probably the biggest obstacle to effective use of large volumes of data. Inconsistent data formats, inconsistent non-aligned data structures, and semantics represent significant challenges that, if not overdue, can lead to project failure.

• Speed: has relation to the response time for a given request. With the real-time communications, it has been increasingly more possible to achieve a bigger velocity to exchange data and information (RIBEIRO, 2014). Mcafee and Brynjolfsson (2012) argue that for many applications, the speed of data creation is even more important than the volume of data, since real-time or near-real-time information makes it possible for a company to be more agile than its competitors.

• Veracity: is related to the quality or reliability of the data. A definition of this characteristic was presented by Erl, Khattak and Buhler (2016), who state that the data need to be evaluated for reliability, which may require specific activities to identify those that do not meet this premise and remove them from the datasets . A data considered of poor quality is one that can not be converted into information and therefore has no value, whereas quality data have value and generate significant information.

• Value: it is related to the return on investment and is the result of the combination of the aspects mentioned above. This feature is intuitively related to the veracity characteristic, since the higher the quality of the data, the more value it has for the business (KAISLER et al., 2013). Likewise, value and velocity can be considered to have inverse relationships, since the longer it takes for the data to be transformed into relevant information, the less value it will have for the business, since obsolete results hamper quality and speed in decision making.

• Variability. refers to data that is constantly changing, such as weather data. Gandomi and Haider (2015), affirm that the variability is related to the variation in data flows.

With the characteristics presented it is possible to determine the notion of the term big data. However, experts believe that there is no need for the existence of all factors at the same time for an environment to be considered big data, since there are cases where there is a greater emphasis for one or another characteristic, and in other cases they are not identified in its entirety

4 METHODOLOGIC PROCEDURES

The query to the Web of Science was made in July 2017 and the string: ("data quality" and "big data") was used in the "Title" search field. Regarding the type of document, only articles and proceedings paper indexed by the database were excluded, excluding documents characterized by book review, editorial material and review among others of less occurrence until the year 2016, resulting in 23 articles on the terms searched. After the identification of the publications, searches were made through the institutional access of the Federal University of Santa Catarina (UFSC), through the full texts of the articles and of the 23 studies found, 6 of which were not freely available for institutional access. For this reason, their summaries will not be presented in this article, but they will be part of the quantitative analysis presented in the next section.

It is worth mentioning that at first, it was defined that the search would include the field of research Topic, which includes the query to the title, to the abstract and the keywords, which resulted in a total of 158 publications, for the same conditions of search. However, after a preliminary analysis in the titles of the publications, it was verified that the great majority did not focus specifically on data quality in big data environments, so, in order to contemplate only the studies related to these two areas, it was decided for using only the Title field.

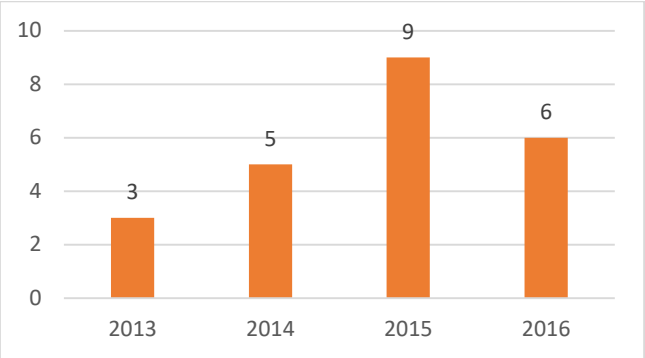
Another issue to be taken into account was the option to use only the term related to data quality and not to make use of the term quality of information. Some authors do not distinguish between the concepts of data quality and information quality, treating both as having the same meaning. However, it is believed that the data are a set of values or occurrences that after processing generate the information (SOMASUNDARAM;

SHRIVASTAVA, 2011), so in the present study, it will be accepted that given and information have different meanings and for this survey will be considered only publications that reference data quality.

The Web of Science database was chosen because it is considered an interdisciplinary database and an important index of scientific journals, enabling quantitative analysis of the results obtained through the consultations (CAPES ' JOURNAL PORTAL/MEC, 2014).

5 RESULTS

The research conducted in the Web of Science on 05/07/2017 for the terms "data quality" and "big data" retrieved a total of 17 proceedings papers and 6 articles, totaling 23 publications. As shown in Graph 1, publications on the subject started in 2013 with 3 publications this year, with the largest number of published studies, 9 in total, in 2015.



Graph 1: Number of publications on data quality in big data
Source: WoS 2017

Most of the studies originated in the USA, with 8 of the 23 publications, Spain and Italy accounting for 2 each and countries like Brazil, Germany, United Arab Emirates, Switzerland, France, Canada, Holland, South Korea, Scotland, Saudi Arabia, China , Australia and Mauritius obtained 1 publication on the subject.

Table 1 presents the title, the author (s), the year of all the publications found and the research contexts. Of the 23 studies, 6 do not have their complete texts available through the institutional access of UFSC - Federal University of Santa Catarina are: 3, 4, 7, 9, 12 and 16, and for this reason their summaries will not be presented.

Chart 1: Publications on data quality and big data.

	TÍTULO	CONTEXTO
1	Pay-as-you-go data quality improvement for medical centers (ENDLER; BAUMGAERTEL; LENZ, 2013)	Financial data management in the health field
2	Information governance, big data and data quality (FREITAS *et al*., 2013)	Information managing
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)	Supply chain data quality management
6	Data quality management, data usage experience and acquisition intention of big data analytics (KWON; LEE; SHIN, 2014)	Data quality management
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)	Data quality management
9	Framework for social media big data quality analysis (AL-HAJJAR, 2015)	-
10	From data quality to big data quality (BATINI *et al*., 2015)	Data quality management
11	Big data, big data quality problem (BECKER; MCMULLEN; KING, 2015)	Data quality management
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)	Data quality management
14	Taking a 'big data' approach to data quality in a citizen science project (KELLING *et al*., 2015)	Information management in Bird monitoring
15	Data quality issues in big data (RAO; GUDIVADA; RAGHAVAN, 2015)	Biologic data quality management
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)	Smart cities data quality management
18	Big data quality: a roadmap for open data (CIANCARINI, 2016)	Open data quality management
19	Data quality: experiences and lessons from operationalizing big data (GANAPATHI, 2016)	Data quality management
20	Antecedents of big data quality an empirical examination in financial service organizations (HARYADI *et al*., 2016)	Data quality management
21	A data quality in use model for big data (MERINO *et al*., 2016)	Data quality management
22	Big data quality - whose problem is it? (SADIQ;	Data quality management

	PAPOTTI, 2016)	
23	Big data quality: a quality dimensions evaluation (TALEB *et al*., 2016)	Data quality management

Source: Web of Science, 2017

The following will present the objectives and a brief description of the publications found.

The study by Hazen et al. (2014) published in the International Journal of Production Economics in 2014 aims to propose a model from the perspective of data science, predictive analysis and big data for continuous improvement in data production, focusing on supply chain management. The proposed model is applied in a case study, as thus are presented its results.

From an empirical research, Kwon, Lee and Shin (2014) presents a mapping of the issues that lead organizations to adopt big data scenarios. In the study, the methodology used for the application of the research is detailed and the results obtained are described, showing that mainly external factors such as the global economic situation, market pressure to offer better services or products and new business opportunities are determining factors for adoption of new strategies such as big data and big data analytics. This article was published in the International Journal of Information Management. Another study with a very specific domain of application was published by Kelling et al. (2015) in AMBIO. The article presents a proposal for the use of big given in a bird monitoring project called eBird, and aims to increase the quality of the data generated and made available to the researchers of the project.

The article published in Future Generation Computer Systems - The International Journal of Science by Merino et al. (2016) considers data quality in big data from the user's perspective and proposes a model called "3As Data Quality-in-Use model." The same is based on ISO / IEC 25012 and ISO / IEC 25024, and researchers suggest that the set of dimensions that make up the model are divided into three groups: contextual adequacy, adequacy, time and operational suitability.

Some publications are focused on the identification of problems related to data quality in big data. Batini et al. (2015) report in the Journal of Database Management an analysis of data quality problems and propose a conceptual framework to monitor the quality according to three points considered by them relevant in big data environments, they are: the type of the data, its origin and the domain of the application.

Becker, Mcmullen, and King (2105) and Ganapathi and Chen (2016) presented studies involving a survey of the problems related to data quality proposing a set of solutions for them. The paper by Rao, Gudivada and Raghavan (2015), addresses the difficulties identified during the integration and sharing of data, focusing on biological data. A similar study to the previous one is presented in Haryadi et al. (2016), but in this publication the objective was to

identify the problems related to data quality in a big date in the context of financial institutions, where 11 data quality dimensions to be considered in the domain in question were defined and an evaluation was performed in 3 institutions , describing the methodology applied, as well as the results obtained. The last four papers were presented at the IEEE International Conference on Big Data in 2015 and 2016.

Also in health and aiming at continuous improvement, Endler, Baumgaertel and Lenz (2013) present in the Conference on eHealth - Health Informatics Meets eHealth 2013, a model based on the TDQM to be adopted in monitoring the quality of financial data health centers when, at the time the proposal was published, the model was still in the validation phase, with no effective results.

In the context of smart cities, Yang, Silva and Picard (2015) propose a framework that was presented at the International Workshop on Computational Intelligence for Multimedia Understanding in 2015 to calculate generic quality indicators in the data flows of smart meters based in complex event processing technology Complex Event Processing (CEP), at the time of publication the proposed framework was already in use in France.

The open data quality (Open Data) is discussed in the study of Ciancarini, Poggi and Russo (2016) and presented at the 2nd IEEE International Conference on Big Data Computing Service and Applications. According to the authors, guaranteeing the quality of the open data is one of the great challenges of this movement and the main objective of the study is to perform an analysis of the data made available by institutions linked to the Italian government and the model used in the evaluation is the ISO/IEC 25012.

The study by Taleb et al. (2016), suggests that the big data quality assessment be done by sampling. Researchers have developed an algorithm for quality assessment that is applied in a case study, and argue that using this approach is efficient as it reduces the time and computing resources involved. The study was presented at the 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 in 2016.

A review of the literature on data quality management in different contexts is presented by Juddoo (2015) at the International Conference on Computing, Communication and Security, and Freitas et al. (2013) present a brief discussion of "information governance". The publication brings concepts about big data and data quality in a succinct and brief way, the study was presented at the IEEE 16th International Conference on Computational Science and Engineering.

Finally, the publications of Saha and Srivastava (2014) and Sadiq and Papotti (2016) do not deal specifically with a study, but rather with the programming of a lecture on the

challenges of data quality management in big data and of a table Roundtable on Big Data Quality Responsibilities respectively, both presented at the International Conference on Data Engineering in 2014 and 2016.

4 FINAL THOUGHTS

As discussed in this article, it can be concluded that discussions related to data quality in big data environments are somewhat recent, since the first publication in the Web of Science database was in the year 2013. Another relevant question are the contexts where the surveys took place, it is noticed that most publications present studies related to the management of data quality in big data environments in a generic way, without specifics in the field of application, since few researches were directed to a specific context , such as financial. Regarding the place of publications, it is observed that the journals or events where they occurred have a tendency to be related to the area of the computer, which is understandable because the big data theme has a strong relation with this area.

Based on the analysis of the results obtained, it is possible to affirm the existence of research opportunities on the topic of data quality in big data within Information Science, which could contribute to the application of data and information concepts that go beyond technology, such as management and flow of them. Because they understand that the data have characteristics and needs inherent to the context to which they belong, other suggestions for future work would be researches that contemplate models to be applied in specific areas of knowledge, as well as the realization of surveys like this in other databases, which may be added to the publications presented in this article.

REFERENCES

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 Computing Surveys**, n.3, v.41, 2009, p. 1-52. Disponível em: <http://dl.acm.org/citation.cfm?id=1541883>. Acesso em: 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>. Acesso em: 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>. Acesso em: 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/>. Acesso em: 15 jun. 2017.

CIANCARINI, Paolo; POGGI, Francesco; RUSSO, Daniel. Big Data Quality: a Roadmap for Open Data. *In*: 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 >. Acesso em: 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>. Acesso em: 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 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, Patricia Kuzmenko; 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>. Acesso em: 21 abr. 2017.

GANAPATHI, Archana; CHEN, Yanpei. Data Quality: Experiences and Lessons from Operationalizing Big Data. *In:* 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>. Acesso em: 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. Disponível em: <http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>. Acesso em: 21 abr. 2017.

HARYADI, Adiska Fardani *et al.* Antecedents of Big Data Quality An Empirical Examination in Financial Service Organizations. *In:* 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>. Acesso em: 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.

Disponível em: <http://www-sciencedirect-com.ez46.periodicos.capes.gov.br/science/article/pii/S0925527314001339?via%3Dihub>. Acesso em: 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>. Acesso em: 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. Acesso em: 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/>. Acesso em: 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. Disponível em: http://www-sciencedirect-com.ez46.periodicos.capes.gov.br/science/article/pii/S0268401214000127?via%3Dihub. Acesso em: 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. Acesso em: 7 jul. 2017.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data. The management revolution. **Harvard Buiness Review**, 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>. Acesso em: 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>. Acesso em: 19 abr. 2017.

SADIQ, Shazia; PAPOTTI, Paolo. Big Data Quality - Whose problem is it? *In:* 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>. Acesso em: 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>. Acesso em: 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. Disponível em:

<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>. Acesso em: 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.

DOI 10.20396/rdbci.v16i1.8650412