ARTICLE

GENDER AND NUMBERS:
using data from International Women's Day coverage on the sites of three major Brazilian newspapers

ANA CAROLINA ARAÚJO
Universidade Federal da Bahia, Salvador – BA, Brasil
ORCID: 0000-0002-3758-0083

DOI: 10.25200/BJR.v15n1.2019.1061

ABSTRACT – Over the last decade, data journalism has evolved alongside information and communication technologies, spreading more and more into professional journalism. The purpose of this article is to understand if and how traditional press has reacted to the new technological tools available for reporting and the increasing availability of open data, whether public or private. To investigate these issues, articles published on the websites of three major Brazilian newspapers – *Folha de S. Paulo, O Globo* and *O Estado de São Paulo* – were analyzed for their coverage of International Women's Day in 2017. The results indicate that there is still a lack of data in most of the material, that the use of digital visualization techniques is still in its early stages and that there is no raw data supply in news production. On the other hand, the use of data and statistics without reference or source checking is quite noticeable and frequently occurs in the corpus.

Key words: Data journalism. Data-driven journalism. Digital journalism. International Women's Day.

GÊNERO E NÚMEROS:
o uso de dados na cobertura do Dia Internacional da Mulher nos sites de três grandes jornais brasileiros

RESUMO – Na última década, o jornalismo de dados evoluiu ao lado das tecnologias de informação e comunicação, espraiando-se cada vez mais pelo ambiente do jornalismo profissional. A proposta deste artigo é compreender se e como o jornalismo impresso tradicional tem reagido às novas ferramentas tecnológicas disponíveis para a prática da reportagem e à crescente disponibilidade de dados abertos, sejam eles de origem pública ou privada. Para investigar estas questões, foram utilizadas como material de análise reportagens publicadas nos sites de três grandes jornais brasileiros – *Folha de S. Paulo, O Globo* e *O Estado de São Paulo* – durante a cobertura do Dia Internacional da Mulher em 2017. Os resultados indicam que a presença dos dados ainda ocorre na menor parte do material analisado, que o uso das técnicas digitais de visualização é incipiente e a oferta de dados brutos no processo de produção da notícia não existe. Por outro lado, foi possível perceber que o uso de dados e estatísticas sem indicação ou checagem de fontes é bastante frequente no corpus estudado.

Palavras-chave: Jornalismo de Dados. Jornalismo Guiado por Dados. Jornalismo Digital. Dia Internacional da Mulher.

GÉNERO Y NÚMEROS: el uso de datos en la cobertura del Día Internacional de la Mujer en los sitios de tres grandes periódicos brasileños

RESUMEN – En la última década, el periodismo de datos creció al lado de las tecnologías de información y comunicación, tomando el campo del periodismo profesional. La propuesta de este documento es comprender si y cómo el periodismo impreso tradicional ha reaccionado a las nuevas herramientas tecnológicas disponibles para la práctica del reportaje y la creciente disponibilidad de datos abiertos, ya sean de origen público o privado. Para esta investigación se analizarán reportajes publicados en los sitios de tres grandes periódicos brasileños – *Folha de S. Paulo, O Globo* y *O Estado de São Paulo* – durante la cobertura del Día Internacional de la Mujer en 2017. Los resultados indicaron bajo uso de datos, uso incipiente de visualizaciones digitales y ninguna oferta de datos brutos. El uso de datos y estadísticas sin referencias o chequeo de las fuentes fueron frecuentes.

Palabras clave: Periodismo de datos. Periodismo guiado por datos. Periodismo digital. Dia Internacional de la Mujer.

1 Introduction

It was November of 2008 when computer scientist Joe Hellerstein predicted in one of his blogs on technology innovation that humanity is currently living in the initial stage of the fourth industrial revolution: the “industrial revolution of data”. He believed the process was in its initial stages because most of the digital information available at that time was ‘handmade’ in the form of web pages and data entered into forms. He states we are at the beginning of the mass production of digital data with software logs, GPS, and other automatic processes, while at the same time the capacity of hard drives and flash drives are increasing at a relatively low cost. The last stage of this revolution is the ‘commoditization’ of data analysis software and the overall generalization of processes which are available to a broad class of users, leading to a volume of data that will quickly surpass the collective production of content writers throughout the world. If this is the direction in which things are going, we have already reached a turning point. We are living in the age of big data[1], an immeasurable amount of data is generated every second around the world, there are hard drives for domestic use with

10-terabyte capacities (not even mentioning cloud drives), and all this data becomes commodities.

In 2010, two years after Hellerstein's analysis, the inventor of the world wide web, Tim Berners-Lee[2], claimed, in a conference at GOV 2.0 Expo, that data-driven journalism (DDJ) was the future and that it would not be enough for journalists to just have general data skills, they would need to know how to further develop these skills, how to analyze and how to determine what is of interest (Arthur, 2010). Since then, what used to be called data journalism (or data-driven journalism (DDJ) as some writers prefer to call it, we will discuss these reasons later) started to find the connections between news and technology interesting.

This study is based on the issues worked on by Hellerstein and Berners-Lee: the intersection between mass data and those interested in it and the possibilities of using this data for producing news. We use journalistic production as the basis for analyzing the use of data in news narratives.

More than just trying to label something as data journalism or not, we intend to figure out if the data is put towards developing a further understanding of the issue, and if the digitalization of contemporary life is being ignored in the news production process. To do this we analyze the data used for verifying, the level of analysis, the origin of the data, the access to data sources and available customized visualizations.

Since this is an exploratory study intended to identify the problem and investigate the analysis methodology, and then use this information for a larger study, we opted for a small *corpus*. We referenced coverage of International Women's Day and empirically observed that every year, one day before the date is celebrated, coverage increases, and public data and studies on the issue are released. Our object of analysis, factual or not, was the March 8th coverage from the three largest newspaper sites in Brazil[3]: *Folha de S. Paulo*, *O Globo* and *Estado de S. Paulo* (which we shall refer to as *Estadão*, the same name given to its internet news portal).

It is important to highlight that this study could have been conducted using a wide array of journalistic material on the issue. Nevertheless, we understand that one of the potential benefits of using data in journalism is the possibility of treating certain issues as a matter of public agenda, and not anecdotal, so we decided that women's issues would be an appropriate base for this study.

DOI: 10.25200/BJR.v15n1.2019.1061

2 The potential of open data

Having the technology and tools available for producing reports supported by data in agendas, verification, information architecture, visualization and publication, does not necessarily mean these resources are actually used. But if there is data available it should be used appropriately, resulting in good quality journalistic material.

The movement of open data goes beyond the field of media. In fact, it adapted to the global market as a device for control and accountability, reaching all levels of government and consolidated the Information Age (Castells, 1999). It is a threefold process: opening, participation and collaboration (White House, 2009, paragraph 4) where transparency, sharing and collective work help to realize the potential of these data sets. Providing access to this data and making it understandable is a form of democratization. We need to remember that the focus of the movement is to publish public data or data generated from public funding as the use of public resources should benefit all of society. Where data from private companies are concerned it is necessary to consider the ownership of its creators (Kitchin, 2014, p. 45).

In Brazil, the legal framework which permits open public data includes supplemental laws 101 (2000) – public finance rules on fiscal management – and 131 (2009) – adds tax enforcement information to law 101 (2000). It also includes federal law 12.527 (2011) – establishes constitutional law on public access to public information and normative ruling 4 (2012) from the Secretary of Logistics and Information Technology (SLTI), Ministry of Planning, which instituted the National Infrastructure of Open Data (INDA). This structure strengthens the production, dissemination and consumption of data. Through its own presence in society and use of public services, the population generates massive amounts of data which are then collected, stored and structured by governments, who theoretically share this information. In the example analyzed in this paper, journalists benefit from open data and are able to help the public understand and interpret this data, translating the raw data into concrete information which is made accessible to non-specialized groups.

Federal law 12.527, also known as the Information Access Law (LAI), moved the active transparency defined in the 1988 Federal Constitution to the practice field, obligating public organs to publish data and information on the institution and services it provides. Requiring this data to be published in open digital formats and not just for owners was the first step towards having open data in Brazil. The mark of this change came with Brazil's inclusion in the Open Government Partnership in 2011, one of the eight founding countries which today welcomes 59 others, all committed to distributing and encouraging transparency in governments, in access to public information and in social participation. This is how Brazil was included in the open government initiative which is predominant around the world (Prado, Ribeiro & Diniz, 2012, p. 16).

These complex and extensive data bases open the doors for the possibility of bigger journalistic endeavors, for discovering corruption scandals, for understanding outbreaks of endemic diseases, etc. It is important to not forget that the management of these data bases also holds a large potential for daily coverage in journalistic media such as municipal public transport, basic sanitation and government spending. It is possible that the higher potential of using data in journalism is not in specialized teams but in its daily use in newsrooms as a form of investigating and contextualizing issues of public interest. Even still, the question posed above remains: has this been happening?

3 The evolution of data use in journalism

The discussion held in this paper requires a questioning of the concept of data journalism itself. If we consider the simple use of data for producing news material as a central characteristic, our discussion will be empty as practically any journalistic practice could be classified in this way. The use of numerical data in journalism began at the end of the nineteenth century as a result of a larger change which was the "substitution of religious paradigms for scientific explanations [...] [that] acquire the status of 'revealing reality'" (Sponholz, 2009, p. 58, our translation). In all the social sciences, collecting, classifying and analyzing data has become an almost indisputable path towards obtaining reality, the truth.

Using data to improve on journalistic procedures is nothing new. It goes back to the twentieth century when Phillip Meyer's concept of precision journalism was established (1973), and later, computer assisted reporting – CAR (Cox, 2000; Gray, Bounegru &

DOI: 10.25200/BJR.v15n1.2019.1061

Chambers, 2012; Coddington, 2014). Both concepts were introduced to the world during the 1980s, and arrived in Brazil the following decade. Both of these practices were established in order to be able to collect, process and extract meaning from available data sets. It represented a transition from a few private databases, most of which were inaccessible, to an infinite amount of open and accessible databases, brought on by the global search for transparency and openness, especially after the year 2000 with the onset of the web. This abundance of databases brought the strong expectation, particularly in journalistic companies, that the production process would now be based on data, and this was the new way to avoid inaccurate narrative journalism (Alexandre, 2014, p. 48; Charbonneaux & Gkouskou-Giannakou, 2015, p. 268). This enthusiasm with data led traditional journalistic companies to look for professionals who could work with data and to invest in training their existing staff (Bertocchi, 2016; Grandin, 2014; Renó & Renó, 2016). At the same time, and as a result of the changing market, there were new products working with data in reporting that emerged, such as *ProPublica* in the United States, and the *Nexo Jornal* and *Agência Pública* in Brazil.

According to Megan Knight (2015, p. 55), the expression "data journalism" first appeared in a text written by Simon Rogers for the English newspaper *The Guardian.* In the blog "Guardian Insider", he wrote about using an application to generate a visualization of raw data in the form of dynamic maps. While explaining how journalists and developers were working together to offer a new reading of the data, he said: "This is data journalism" (Rogers, 2008, paragraph 7). In a video published in 2013 (Rogers, 2013), the author claims that in 1821 *The Guardian* newspaper leaked cost spreadsheets for local schools, thereby claiming that this newspaper invented Data Journalism. This was not the first time that newsrooms had worked with data, but the form that was used here was definitely new.

A review of literature shows us that the first registered use of data in journalism occurred in 1952 when TV broadcaster CBS used computer programs to try and predict the results of the United States presidential elections (Gray, Bounegru & Chambers, 2012, p. 9). However, the first systematized register of data use in journalism came later, in 1967, with Philip Meyer and his decision to apply methods used in social sciences to improve the quality of journalistic verification, a work that resulted in publishing his book *Precision Journalism* in 1973. Looking at objectivity from a positivist viewpoint, Meyer understood that the scientific method would be capable of producing better journalism than what had been published in press releases and statements from sources. Years later, he put his concept into perspective, and researched new techniques for analyzing and processing data in *The New Precision Journalism* (1991). In a statement to Bounegru in 2012, more than 40 years after releasing precision journalism, Meyer came across a little skeptical when he wrote: "When information was scarce, most of our efforts were devoted to hunting and gathering, now the information is abundant, processing is more important" (Gray, Bounegru & Chambers, 2012, p. 6).

In the 1990s, the study on the use and processing of quantitative data in journalism was combined under the term computer assisted reporting (CAR). There are indicators that Meyer's work has been one of the main driving forces, both inside and outside of academia, of consolidating the term and practice of CAR, defined by Cox (2000, p. 3) as a broad concept which encompasses the different uses of computers in producing news, including online research and databases. Literature on the use and improvement of computers in newsrooms talks of how the practice of journalism merges with the popularization of microcomputers and digital files, as well as the internet (Machado, 2005, p. 301; Flew, Spurgeon, Daniel & Swift, 2012, p. 157, Crucianelli, 2013; Coddington, 2014, p. 2; Cushion, Lewis & Callaghan, 2016, p. 3).

Computer use and web tools, as well as open data, grew immeasurably in the second half of the twentieth century, and exponentially in the 2000s (Kitchin, 2014, p. 79; Zuiderwijk, Janssen & Dwivedi, 2015, p. 431).

In journalism, this breakthrough occurred in 2010 when thousands of Afghanistan and Iraq war logs were leaked by the website WikiLeaks[4]. The website's founder, Julien Assange, and a team from *The Guardian*, published thousands of high-quality data, although disappointingly complex. The solution was to develop a personalized web browser using keywords to get concrete access to the information contained in big data (Rogers, 2014; Hewett, 2015).

Since the "industrial data revolution" foreseen by Hellerstein in 2008 became a reality (in which she describes that the volume of data produced by automatic and intelligent processes would undoubtedly surpass that of 'manufactured' production), her journalism has been quickly appropriated. Therefore, deciding

DOI: 10.25200/BJR.v15n1.2019.1061

whether the data became a new category for sources or a new set of tools for journalism is important here. For data journalism, one thing is not separate from the other, and the role of professionals in this field is to create connections between data which is initially incomprehensible (Bradshaw & Rohumaa, 2011; Bradshaw, 2014). Now citizens do not need the press in order to access data, but the same cannot always be said about its meaning. In this scenario, the institute of "scoops" as a symbol of excellence in a professional field has disappeared. The debate isn't necessarily about who was first to release the news, instead, it's increasingly about who can make connections that the others are not capable of doing, who sees what the others do not see through the use of journalistic competences and a mastery of the technique.

In order to further our discussion we decided to briefly analyze the terms and concepts used to define the set of practices that, up until now, we have referred to as data journalism. Even though some journalists and professors in the field claim that data journalism (DJ) is just a new name for computer assisted reporting (CAR), we believe there are differences in terms of objectives and the final product. This final product can be thought of as an evolution of Precision Journalism (Meyer, 1973), it is largely supported by methodologies in social sciences (surveys, content analysis, statistical analysis, etc.) in order to build up investigative reporting (Coddington, 2014, p. 4). The news product in DJ originates from the database itself which is equipped with an abundance of information for building an analytic narrative. In the words of Stray (2011, p. 19), it is the act of obtaining, selecting and reporting data of public interest. More than just refining the collection of information, the focus is on data being open to the public, including knowledge from different fields, especially journalism, programming and design.

In an attempt to step away from the idea that all journalism comes "from data", Parasie & Dagiral (2012, p. 854) uses the term "Data Driven Journalism" (DDJ) to describe a more updated practice, capable of improving the democratic contribution of journalism in three ways: strengthening journalistic objectivity, maintaining accountability in governments through journalistic companies, and increasing citizens' political participation through its production and data analysis. The term is a reference to Träsel who, in his doctoral thesis, described DDJ as "the application of computers and knowledge from the social sciences in the interpretation of data, with the aim of expanding the role of the press as a defender of public interest" (Träsel, 2014, p. 90, our translation).

Along these same lines we have the concept of computational journalism, as proposed by Hamilton & Turner (2009, p. 2), which is a combination of algorithms, data and knowledge from the social sciences used to supplement the accountability function of journalism, uniting data journalism and computer-assisted reporting. A perspective which seems to us to be more appropriate is that of Diakopoulos (2011, paragraph 2), he describes it as a technological component which is guided by journalism, applying computational thought to collection practices, to producing meaning and to presenting information using automated processes.

The lack of a clear delimitation between the concepts presented above is noticeable. Conversely, some writers believe the concepts overlap one another in many ways (Coddington, 2014; Lima Júnior, 2012; Mancini & Vasconcellos, 2016). In order to create some harmony between these concepts, we suggest that these data journalism practices fit into extensions of the paradigm of digital journalism in databases (DJDB), a theoretical model formulated by Barbosa & Torres (2013) with the purpose of better understanding the function of databases in contemporary journalism. This model expands in consecutive appropriations, keeping in mind that current databases are integral journalistic activities and that they are singular agents in the convergence process of communication media, solidifying its prominence (p. 153).

Therefore, we understand data driven journalism to be one of the components, one of the aspects of PDJD (Barbosa & Torres, 2013), since it covers the range of its concept, summarized as:

> [...] the model that has databases for defining structure and organization, as well as the composition and presentation of journalistic content, according to functions and specific categories, which also allow for the creation, the maintenance, the updating, the availability, the publication and the circulation of dynamic cyber mediums across multiplatforms. (Barbosa & Torres, 2013, p. 154, our translation).

The analysis methodology used here was based on some of the works previously mentioned in this paper which we shall outline below.

DOI: 10.25200/BJR.v15n1.2019.1061

4 Methodology

The growth of available information on the internet has been addressed by different writers with the hope it can bring about social and political change for democratic states (Marques, 2008, p. 50; Pinho, 2008, p. 476; Rothberg, 2008, p. 151; Silva, 2009, p. 41; Bragatto, 2011, p. 133), this approach repeatedly being linked to upholding civil, political and social rights as mediators of journalism. Digital media and the availability and use of increasingly robust and intercommunicative databases (big data and open data) as predicted by Berners-Lee in 2010 have expanded the field of journalistic work, providing a large amount of data for continued discussions on issues of public interest, which include health, education, public safety, economic development, ethnic diversity, religion and sexual health, etc.

The aim of this paper is to understand if and how this apparatus of informative sources is being used in journalism. To analyze this we chose agendas connected to women's issues, in particular the increasingly recurrent number of publications on this issues leading up to March 8th, International Women's Day. The *corpus* of this study includes material on women published on March 7, 8 and 9 of 2017 on three of the larger journalistic sites in Brazil: *Folha de S. Paulo, O Globo* and *Estadão*. The data was taken from the Institute for Verifying Communication (IVC)[5], which investigates the circulation of each newspaper, including print copies and digital access. It is of note that at the time the data was collected the three publications chosen for this paper were first, third and fourth in IVC ranking, respectively. In second place was Super Notícia, a free journal in Minas Gerais put out by the group O Tempo which does not have any search mechanisms on its site, and for this reason it could not be included in this study.

In total there were 129 articles to examine, register and discuss: 38 from *Estadão*, 41 from *Folha de S. Paulo* and 49 from *O Globo*. In order to refine the data we only considered material which was produced by the media newsrooms and which represented their editorial lines and organizational culture, disregarding material from news agencies and blogs hosted on the investigated domains.

The analysis combines variables tested in three different studies, as well as four criteria proposed for the first time in this study, all combined into one electronic spreadsheet. We searched the digital archives of each of the publications and selected material which contained the word woman in the title or in the body of the text and was published during the designated time period. All content was downloaded individually in PDF files to ensure that it could be used throughout the whole study regardless of any later changes that may have occurred on the news portals we visited. After collection, each piece of material was closely examined in order to complete the instrument.

The first set of variables is an adaptation of the typology proposed by Megan Knight (2015, p. 59) for analyzing data published in British newspapers in order to observe supports and formats of data journalism as a *praxis* of traditional media (N=106). She classified the material into the following categories: media groups, sections, data usage category and origin of data.

Added to these categories were categories from Cushion, Lewis & Callaghan (2016, p. 7), who analyzed an extensive database for a study commissioned by BBC Trust on the relation between journalism and data in journalistic coverage in the United Kingdom, including radio and television programs and online media (N=4.285). The following variables were then selected: data usage, data source and problematization of data.

Three more categories were adopted in order to improve the analysis, this time they were taken from the Paradigm DJDB (Barbosa & Torres, 2013, p. 155): interrelationship/*hyperlinkage*, visualization and convergence. Recognizing the specific nature of this journalistic practice, we found it appropriate to include three categories which were not included in the bibliography used for this study: diversity of sources (one, two, three or more data sources), access to numerical data sources (cited data sources, uncited data sources, cited data sources and offers an access link), and making the raw data available. Each analyzed variable is described below as well as the operators who classified them:

DOI: 10.25200/BJR.v15n1.2019.1061

Table 1 – Categories of analytical operators

Category	**Operators**
Media – company responsible for the report	1. Folha de S. Paulo 2. O Globo 3. Estadão
Section – thematic context of report	1. Brazil 2. City 3. Behavior 4. Culture 5. Economy 6. Sports 7. Infographs 8. International 9. Women 10. Opinion 11. Politics 12. Health 13. Society
Data usage – indicate the presence of any numerical data and its depth level	1. None 2. Vague reference, no empirical support 3. Brief but clear reference 4. Detailed reference with context
Type of data usage – once verified, classify the data usage	1. Static map 2. Dynamic map 3. Graph or chart 4. Infograph 5. List 6. Numerical highlight 7. Table 8. Textual analysis of data 9. Timeline 10. Citing data
Diversity of sources – the number of data sources analyzed	1. One source 2. Two sources 3. Three sources 4. Four sources
Access to sources – checking access to original data source	1. Does not cite a data source 2. Cites a data source 3. Cites a data source and offers an access link
Offers raw data – check if raw data used in analysis is offered	1. Yes 2. No
Data sources –who inserted the data into the material	1. Journalist 2. Interviewee

Origin of data – who produced the cited data	1. Government 2. Private or public institutions 3. Unknown source
Problematization of data – check if journalist overlaps the data with contrary information, questioning a point of view	1. Yes 2. No
Inter-relationship/hyperlink inform if there are web pages offering context and if they favor critical analysis	1. No links 2. Internal links (to the site's database) 3. External links (to other databases)
Visualization – indicate the presence of exclusively digital visualization	1. Present 2. Absent
Convergence – taking into account producing convergent content in newsrooms, registering if and how many formats or supports were used in material	1. One support 2. Two supports 3. Three supports 4. None

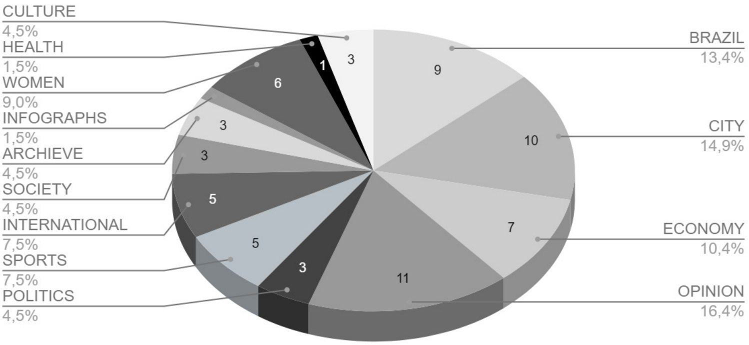
Source: Knight (2015), Cushion (2016), Barbosa (2007) and the author.

The next section will present the results and discussion of the data.

5 Results and Discussion

We found 129 articles on women's related issues between March 7th and 9th, 2017 on the sites of the newspapers *Estadão* (38), *O Globo* (49) and *Folha de S. Paulo* (42). These articles were analyzed according to the criteria laid out in the previous section.

Graph 1– Stories with any data usage, by media



Source: Elaborated by the author.

DOI: 10.25200/BJR.v15n1.2019.1061

As per Graph 1, excluding those articles which do not use data (62 articles or 48.1% of samples), the Opinion section is the one which used the most resources, followed by the sections of City and Brazil. There is some material on health and violence against women in other sections but they did not present any data or statistics. Contrary to what one might believe, the economy section is only in fourth place. The section called Women's Page, which the newspaper's site claimed was created to expand coverage on gender issues, had only four of the newspaper's 42 thematic articles during the study period.

The second category of analysis focused on numerical data and its depth level. According to the graph above, almost half of the material contained brief yet clear information on data, for example:

> "Less than 10% of Congress is made up of women, a number which is much lower than many other Latin American countries and infinitely less than developed countries[...]. "From the article 'Women have less voice, participation and do not have decision-making power'". (Costa, 2017, online).

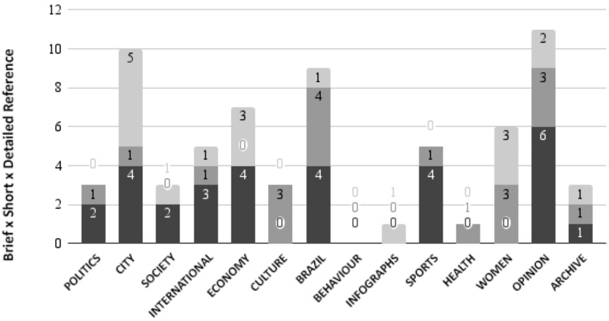
On the other hand, few articles contained detailed data with contextual information to make it reader-friendly, like the sample below:

> "In 2016, one woman was physically abused every 12 minutes in Rio. These numbers, released yesterday by the Military Police and the State Secretary of Safety on International Women's Day, show how severe the situation of gender violence is in the state". (Galdo, 2017, online).

The most interesting finding falls within the middle of these categories. Considering it is a journalistic technique, it is surprising to find out that 28.4% of the material used vague references or did not reference any source or database for their citation, as in the example below:

> "Despite the fact that it is practically illegal for all women to have an abortion, there is a disproportionate number of black, indigenous, poor, and uneducated women who live far from urban centers who are criminalized". (Cardoso, 2017, online).

Graph 2 – Reference deepening, by section

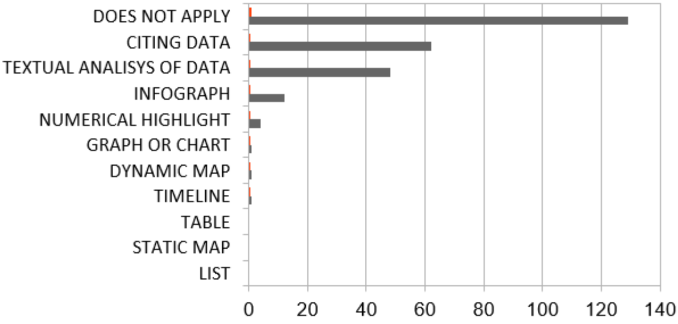


Source: Elaborated by the author.

Looking at this section by section, Brazil and Opinion are the ones where most of the data in the articles was unsupported. The articles which used detailed and contextualized references were found in the sections of City, Economy and Women. Even though the methodology applied here cannot be used to evaluate this, we believe it merits further exploration, including studying the availability of data on certain issues and the production routine of the research teams.

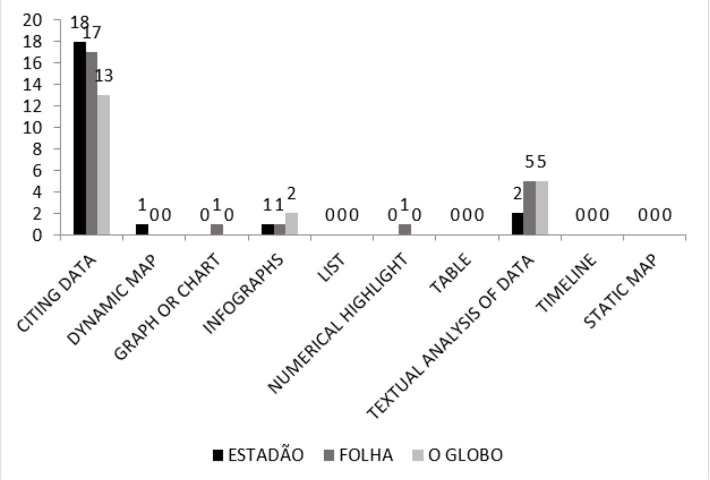
Regarding the types of data usage, Figure 3 shows that more than 70% are simple citations (this is excluding all articles that did not use any citations). This is followed by textual analysis, which we found in 12 articles. Visual resources were barely used, and there were no maps or timelines included. Simple citations were predominant in all the different media vehicles. Of note is the fact that despite having the most articles on the issue, the *Estadão* was the newspaper which used the least amount of data.

Graph 3 – Types of data usage



Source: elaborated by the author.

Graph 4 – Types of data usage, by media

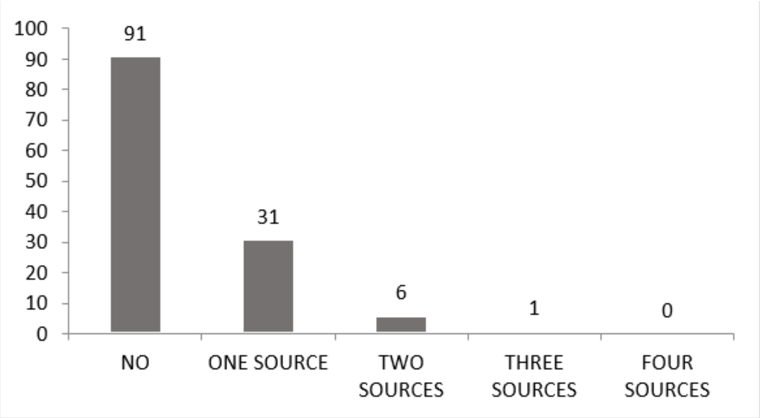


Source: elaborated by the author.

Due to the fact that digital journalism is not affected by the limitation of space as much as print has, we decided to analyze the diversity of data sources used in the texts. When we talk about diversity of sources we are referring to numerical data sources and not to the number of people interviewed in any one report. Our findings showed that there was a very limited amount of diversity, as per Figure 6.

Among the articles using data, more than 80% used just one source, 15% used two and only one article used three data sources. One surprising result was that 24 of the 67 sample articles used numerical references (35.8%). The data was used without any reference to a source, which is in contrast with the basic principles of verification in journalism. Out of all the sample articles, only one on violence against women from *Folha* provided a link to the site where the information was obtained. There was also no raw data on offer in the article, something which the good practice manuals in data journalism recommend (*The Data Journalism Handbook*, 2012).

Graph 5 – Diversity of sources

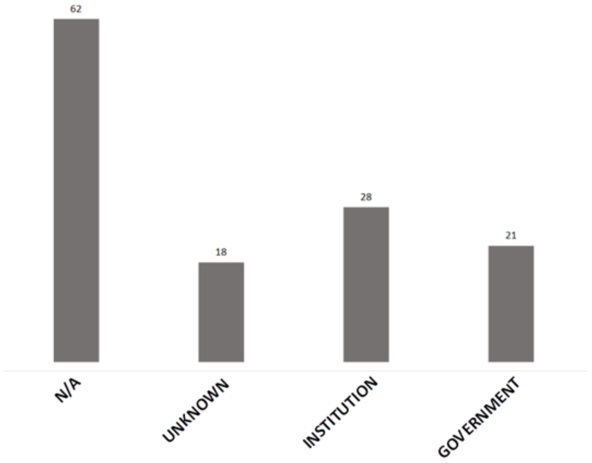


Source: elaborated by the author.

The next step involved analyzing the data sources. Almost twice as many journalists used data as interviewees did. As stated earlier, it is not possible to establish how this result was reached using the methodology in this paper, but we do have a couple conclusions. The first is that it appears that journalists feel a greater need to legitimize discourses through the use of numbers seeing as how their choice of sources tend to possess some kind of authority. The second is that it appears that journalists believe the testimony from their interviewees to be adequate and do not ask them to provide any data to back up their claims.

Continuing with the characteristics of data, the origin of cited information in the writers' texts and from their sources varied between government sources, institutional sources (companies, unions, associations, non-profit organizations, etc.) or unknown sources (where data was cited without indicating who had produced it or published it), as evident in Graph 6. Most of the data cited in texts published by the three newspapers comes from unknown sources; in other words, there is a prevalence of statements which cannot be proven to be true, drawing more attention to the fact that little consideration was made to checking sources during the verification process. There are three explanations for why this might occur. The first is that journalists cite data without revealing their sources, which then makes us wonder if these sources actually exist.

Graph 6 – Origin of the data



Source: elaborated by the author.

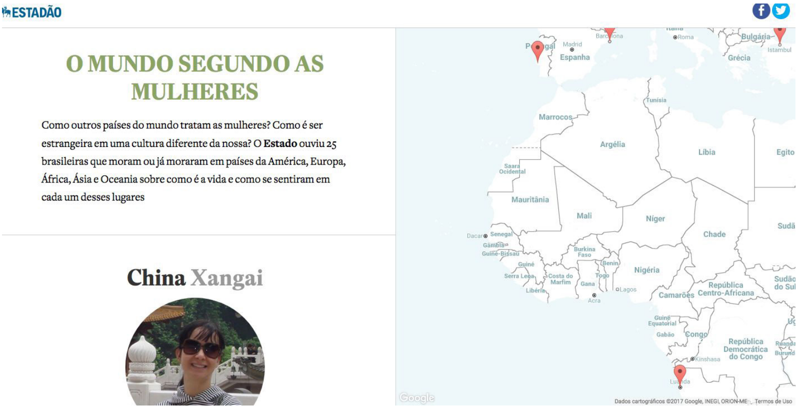
The second is that the interviewees cite data without any references and are not questioned by the reporters about them, and then these same reporters publish the information anyway. This happened 65% of the time that data was used in the sources' statements. Thirdly, articles with this lack of verification make their way through the entire news production process, from editing to printing; and this happened many times.

In terms of the problematization of data, none of the analyzed material presented data used by both the journalist and the source. What we mean is that out of the 17 times that the sources used data in their discourses, regardless of its origin, the journalists never used any additional data to either confirm or refute the interviewees' points of view, and thereby fulfilling their role as journalists by asking questions and offering context and a variety of perspectives about a particular issue or event.

The category of inter-relationship/hyperlinkage was adapted to Barbosa's model (2008) to not only verify the use of the available tool but also to consider that the source could also be used for context. Even though there are limitations involving the production of texts and graphic material for context it would still be possible to lead readers to other approaches around a particular agenda, which according to our results does not happen a lot. 70% of the material we checked did not have any links available, 20% contained links to other articles on the site, and only seven articles (5%) had links to other sites offering more information on the topic. Despite all three newspapers having extensive digital archives which could easily be made available to readers, for some reason they do not offer it.

We mentioned earlier that online journalism somewhat supersedes spatial limitations, well we can also say that it has more data visualization tools to offer, some of them exclusive to digital mediums. This includes not only computer programs that provide a wider range of resources for designers but also a set of free tools for visualization which are easy to use. Even still, there was little importance given to visualization in the samples analyzed in this paper, only six of the 129 articles had them (four infographs, one table and an interactive map in the *Estadão*). Even if we look at just the 67 articles which used data, less than 10% of them used visualizations. From another point of view, only one third of of the 18 articles offered with detailed references and context had visualizations.

Figure 1 – Example of an article with an infographic content



ESTADÃO

O MUNDO SEGUNDO AS MULHERES

Como outros países do mundo tratam as mulheres? Como é ser estrangeira em uma cultura diferente da nossa? O **Estado** ouviu 25 brasileiras que moram ou já moraram em países da América, Europa, África, Ásia e Oceania sobre como é a vida e como se sentiram em cada um desses lugares

China Xangai

Source: Estadão (2017). Retrieved from infograficos.estadao.com.br/cidades/o-mundo-segundo-as-mulheres/

The last category presented here is convergence. This category aimed to analyze the use of multimedia for presenting content. Here we found only five cases that used two forms of media in articles, including text and video. Our analysis did not include articles which had text and photos or text and static visualizations since these alternatives can also be used in print newspapers, and doing so does not constitute them being multimedia.

6 Final Considerations

This paper examines and discusses whether there is any truth behind some of the allegations that defenders of data journalism have made, and tries to understand to what level they are a target of appropriation in traditional media newsrooms. This process is addressed specifically to cases in which the availability of open data may contribute to covering a particular theme (and not a scoop), in this case the theme being women. In the end, there are no signs that data usage has made its way into traditional media newsrooms in Brazil, or that the journalists are actually using the mass amounts of data available to qualify their journalistic products. Nevertheless, it is important to state that, at some level, all the media vehicles analyzed for this paper used the set of practices we know as data-driven journalism.

In light of the existence of a specialized team in the area, and despite having produced less material on the theme under analysis during the study period, the *Estadão* was the site which used data the most, proportionately speaking. In addition, their portal was the only one to offer a more advanced digital visualization, a dynamic map. In general, the other two newspapers used data less frequently. In terms of visualizations, the *Estadão* produced three, while *O Globo* and *Folha* each produced two visualizations despite having published a significantly greater amount of material on women during the analysis period.

The literature consulted in this study confirmed the expectation that most of the data which was utilized came from government sources, and therefore leaned more toward discourse from social actors in positions of power who have privileged access to news production channels. A surprising result, however, was that the use of data from unknown sources was greater than that from government and institutional sources combined, which occurred much more with the journalists than the interviewees. Differently from countries such as England and the United States, material supported by large amounts of data published by governments or other institutions is still at a low in Brazil, which would objectively depict the practice of data-driven journalism. This has a greater impact on communicating issues like health and public safety, for instance. In regards to these issues, the practice of journalism in countries ranked higher on the human development index (HDI) often use this type of source which includes survey results and sets of open data from government.

The idea that enthusiasts of data journalism defend is that the findings that are granted through the Information Access Law and from state information which gets leaked out are important sources for journalistic work, yet this didn't appear to be the case, at least not during the study period and for the issue of this paper. Despite the many events we identified over the last three years on digital mediums, such as the popular phenomenon known as the Spring of Women, everything indicates that agendas on women are still being reported on as they happen, and not as an important issue on the public agenda.

It is also important to remember that even when journalists are interested in data, their ability to use them (which is of urgent necessity according to Berners-Lee in 2013) is still more of an ideal. In other words, even when they use data, most journalists still do so superficially, without looking deeper and contextualizing the information they are presenting.

We are confident that the methodology proposed here, after modifying and testing it many times before ultimately collecting the data, proved to be appropriate and efficient for the purpose of this study. The investigation encouraged debate on the expansion of the journalistic field through the use of large amounts of data as the foundation for discussions on issues of public interest.

As cited earlier, we opted to combine analysis variables from three previous studies, and then added four new criteria we considered necessary to perform a more detailed analysis. This is an innovative proposal seeing as it combines methods together which then provide a new matrix of investigation that each method needed. Thus, the collected material was analyzed according to supports, formats, origin and data sources, technological resources and design, in addition to a more specified look at the practice of journalism and its discursive results when dealing with the problematization of data, diversity of sources and availability of raw data.

Thus, addressing the demands of a critical analysis of discourse in traditional journalistic media, this study offers not only an understanding of what has been said, but also, at more important moments, of what has not been said. For instance, when a journalist or a government authority cites data on violence against women but does not use any objective reference – as

DOI: 10.25200/BJR.v15n1.2019.1061

this study repeatedly discovered – they leave others to interpret about the information source, which might not even exist. Why would such an agent prefer a discourse devoid of meaning when dealing with an issue of such importance? More than just norms and the journalistic way, it is possible to see a few characteristics of professional journalists and their individual attitudes regarding the deontology of journalism.

Considering the results described and discussed in this paper, we understand that despite this analysis dealing with feminine issues published on International Women's Day, the methodology could be applied to study other objects. In terms of women's rights, the methodology would help towards studying other journalistic stories, like the issue of abortion in Brazil, or domestic violence against women, or the gender barriers in the labor market. One suggestion to improve the quality of results would be to broaden the *corpus* to provide more accurate characterization of the trends of data usage in media journalism.

The findings from this study show that the connection between data usage and not only women's issues, but also the democratic development on a whole, help lead to a more promising study. In regards to this article, however, what was demonstrated was that the use of data in traditional newsrooms has not garnered any more attention than any other routine journalistic technique.

NOTES

[1] Mayer-Schönberger & Cukier (2013) state that the explosion in the amount of data started in the sciences with astronomy and genetics in the 2000s, when the term was created. They claim that the initial concept of big data came from the technical incapacity to deal with such a large amount of information. Its importance grew and today these writers describe the term as a reference to things done on a large scale which can no longer be achieved on a small scale such as extracting *insights* or creating forms of value in such a way that they change markets, organizations and the relationship between citizens and governments, etc. Mayer-Schönberger e Cukier, 2013, p.6).

2 British physicist, creator of the *World Wild Web* (WWW or Internet), and founder of the *World Wide Consortium* (W3C), a forum for technological development on the Web, and co-founder of the *Open Data Institute* in London, its mission is to maintain the privacy, freedom and openness of the network.

3 According to the Participation and Position Report in the market section in March 2017 provided by the CVI, a non-profit organization that assesses multiplatform media and provides the market with impartial and detailed data on communication, including web traffic for both desktop computers and smartphones, tablets and applications, as well as circulation, events, inventory and *out of home* media campaigns.

4 WikiLeaks describes itself as an international media organization and collective library, founded by journalist Julian Assange in 2006. It is an organization that specializes in analyzing and publishing vast amounts of official classified or secret data on wars, espionage and corruption, and has already published more than 10 million authenticated documents and analyses.

5 The Participation and Position Report in the market section in March 2017, provided by the Institute for Verifying Communication (IVC).

REFERENCES

Alexandre, I. A. R. (2014). *Jornalismo de Dados: o estado da arte nos jornais generalistas diários em Portugal* (Dissertação de mestrado). Retrieved from Repositório Universidade Nova (hdl.handle.net/10362/13615).

Arthur, C. (2010) Analysing data is the future for journalists, says Tim Berners-Lee. Retrieved from www.theguardian.com/media/2010/nov/22/data-analysis-tim-berners-lee.

Barbosa, S. (2007). *Jornalismo Digital em Base de Dados (JDBD). Um paradigma para produtos jornalísticos digitais dinâmicos* (PhD Dissertation). Retrieved from Repositório Institucional UFBA (repositorio.ufba.br/ri/handle/ri/11299).

Barbosa, S. (2008). Modelo Jornalismo Digital em Base de Dados (JDBD) em interação com a convergência jornalística. *Textual & Visual Media, 1*, pp. 87–106.

Barbosa, S, & Torres, V. (2013). O paradigma "Jornalismo Digital em

DOI: 10.25200/BJR.v15n1.2019.1061

Base de Dados": modos de narrar, formatos e visualizações para conteúdos. *Galáxia, 25*, pp. 152–164. https://dx.doi.org/10.1590/S1982-25532013000200013.

Berners-Lee, T. (2010, Maio 26). *Open, Linked Data for a Global Community* [Video FIle]. Retrieved from www.youtube.com/watch?v=ga1aSJXCFe0.

Bertocchi, D. (2017) *Dos dados aos formatos – Um modelo teórico para o design do sistema narrativo no jornalismo digital.* (Doctoral dissertation). Retrieved from Biblioteca Digital USP. doi: 10.11606/T.27.2014.tde-21092015-122011

Bradshaw, P. & Rohumaa, L. (2011). *The Online Journalism Handbook. Skills to survive in the digital age*. Harlow, England: Pearson.

Bradshaw, P. (2014). *O que é Jornalismo de Dados? Manual de Jornalismo de Dados.* Retrieved from datajournalismhandbook.org/pt/introducao_0.html.

Bragatto, R. C. (2011). Democracia e internet: apontamentos para a sistematização dos estudos da área. *Compolítica*, *2(1)*, pp. 132-163. https://doi.org/10.21878/compolitica.2011.1.2.14

Brasil (2012, Apr. 12). *Instrução Normativa n. 4, de 12 de abril de 2012*: Institui a Infraestrutura Nacional de Dados Abertos – INDA. Brasília. Retrieved from dados.gov.br/pagina/instrucao-normativa-da-inda.

Brasil (2000, May 4). *Lei n. 101, de 04 de maio de 2000*: Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências. Brasília. Retrieved from www.planalto.gov.br/ccivil_03/Leis/LCP/Lcp101.htm.

Brasil (2009, May. 27). *Lei n. 131, de 27 de maio de 2009*: Acrescenta dispositivos à Lei Complementar no 101, de 4 de maio de 2000, que estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências, a fim de determinar a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, do Distrito Federal e dos Municípios. Brasília. Retrieved from www.planalto.gov.br/Ccivil_03/leis/LCP/Lcp131.htm

Brasil (2011, Nov. 18). *Lei n. 12.527, de 18 de novembro de 2011*: Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. Brasília. Retrieved from www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm.

Cardoso, D. (2017, Mar 7). PSOL entra com ação no STF para descriminalizar aborto até 12ª semana. *Estadão.* Retrieved from saude.

estadao.com.br/noticias/geral,psol-entra-com-acao-no-stf-para-descriminalizar-o-aborto-ate-12-semana-de-gestacao,70001690371

Castells, M. (1999). *A era da informação:* economia, sociedade e cultura*. v. 1. A Sociedade em Rede.* São Paulo: Paz e Terra.

Charbonneaux, J.; Gkouskou-Giannakou, P. (2015). O Jornalismo de “Dados”, uma Prática de Investigação? Um olhar sobre os casos alemão e grego. *Brazilian Journalism Research, 11(2)*, pp. 266–291. https://doi.org/10.25200/BJR.v11n2.2015.592

Coddington, M. (2014). Clarifying Journalism’s Quantitative Turn: A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-assisted Reporting. *Digital Journalism, 3(3)*, pp. 331–348. https://doi.org/10.1080/21670811.2014.976400

Costa, M. T. (2017, Mar 8). ‘Mulheres têm menos voz, participação, e não têm poder de decisão’. *O Globo*. Recuperado de oglobo.globo.com/economia/mulheres-tem-menos-voz-participacao-nao-tem-poder-de-decisao-diz-miriam-muller-21028867

Cox, M. (2000). The Development of Computer-Assisted Reporting. Newspaper Division, Association for Education in Journalism and Mass Communication, *Southeast Colloquium, v. 2030*, n. 305, pp. 1-22. Retrieved from citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.631.6220

Crucianelli, S. (Ed.). (2013). *Ferramentas Digitais para Jornalistas 2.0.* Recuperado de knightcenter.utexas.edu/books/FerramentasDigitaispara Jornalistas.pdf.

Cushion, S, Lewis, J. & Callaghan, R. (2016) Data Journalism, Impartiality and Statistical Claims. *Journalism Practice, 11(10)*, pp. 1198-1215. https://doi.org/10.1080/17512786.2016.1256789

Diakopoulos, N. (2011) *A Functional Roadmap for Innovation in Computational Journalism*. Retrieved from www.nickdiakopoulos.com/2011/04/22/a-functional-roadmap-for-innovation-in-computational-journalism/

Flew, T., Spurgeon, C., Daniel, A. & Swift, A. (2012). The Promise of Computational Journalism. *Journalism Practice, 6*(2), pp. 157–171. https://doi.org/10.1080/17512786.2011.616655

Galdo, R. (2017, Mar. 6). Crimes contra mulheres no topo de queixas à PM. *O Globo.* Retrieved from oglobo.globo.com/rio/crimes-contra-mulheres-no-topo-de-queixas-pm-21034906

Gray, J., Bounegru, L. & Chambers, L. (Eds.). (2012) *The Data Journalism Handbook – How Journalists Can Use Data to Improve the News*. Retrieved from datajournalismhandbook.org/pt/index.html.

DOI: 10.25200/BJR.v15n1.2019.1061

Grandin, F. R. (2014) Criação de valor a partir do Jornalismo Guiado por Dados. In *11th World Media Economics and Management Conference.* pp. 1-30. Rio de Janeiro: UFRJ. Retrieved from www.academia.edu/7240997/Criação_de_valor_a_partir_do_Jornalismo_Guiado_por_Dados

Hamilton, J. & Turner, F. (2009) *Accountability through Algorithm: Developing the Field of Computational Journalism*, Stanford, CA, Center For Advanced Study in the Behavioral Sciences Summer Workshop, Universidade de Stanford. Retrieved from web.stanford.edu/~fturner/Hamilton%20Turner%20Acc%20by%20Alg%20Final.pdf

Hellerstein, J. (2008) *The Commoditization of Massive Data Analysis*. Retrieved from radar.oreilly.com/2008/11/the-commoditization-of-massive.html.

Hewett, J. (2015) Learning to teach data journalism: Innovation, influence and constraints. *Journalism, 17* (1*)*, pp. 119-137. https://doi.org/10.1177/1464884915612681

Kitchin, R. (2014). *The Data Revolution – Big Data, Open Data, Data Infrastructures & Their Consequences* **(**R. Rojek) London: SAGE Publication Ltd.

Knight, M. (2015). Data journalism in the UK: a preliminary analysis of form and content. *Journal of Media Practice, 16(1)*, pp. 55–72. https://doi.org/10.1080/14682753.2015.1015801

Lima Júnior, W. T. (2012) Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de dados. *Estudos em Comunicação, 12*, pp. 207–222. Retrieved from www.ec.ubi.pt/ec/12/pdf/EC12-2012Dez-11.pdf

Machado, E. (2005) A Base de Dados como formato no Jornalismo Digital. *Actas do VII Lusocom.* Covilhã: Universidade Beira Interior, pp. 301–310. Retrieved from www.bocc.ubi.pt/pag/machado-elias-base-dados-formato-jornalismo-digital.pdf

Mancini, L. & Vasconcellos, F. (2016). Jornalismo de Dados: conceito e categorias. *Fronteiras – estudos midiáticos, 18(1)*, pp. 69–82.R Retrieved from revistas.unisinos.br/index.php/fronteiras/article/view/fem.2016.181.07

Marques, F. P. J. A. (2008) *Participação política e internet: meios e oportunidades digitais de participação civil na democracia contemporânea, com um estudo do caso do estado brasileiro.* (Doctoral dissertation). Retrieved from Repositório UFBA (https://repositorio.ufba.br/ri/handle/ri/11303)

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data:* A revolution that will transform how we live, work, and think*.* New York: Houghton Mifflin.

Meyer, P. (1973) *Precision Journalism:* A Reporter's Introduction to Social Science Methods. Bloomington: Indiana University Press.

Meyer, P. (1991) *The new precision journalism.* Bloomington, Indiana: Indiana University Press.

Parasie, S. & Dagiral, E. (2013) Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New media & Society, 15(6)*, pp. 853–871. https://doi.org/10.1177/1461444812463345

Pinho, J. A. G. de. (2008). Investigando portais de governo eletrônico de estados no Brasil: muita tecnologia, pouca democracia. *Revista de Administração Pública, 42(3)*, pp. 471–493. https://dx.doi.org/10.1590/S0034-76122008000300003.

Prado, O., Ribeiro, M. & Diniz, E. (2012) Governo eletrônico e transparência: olhar crítico sobre os portais do governo federal brasileiro. In J. A. G. Pinho (Ed.), *Estado, sociedade e interações digitais: expectativas democráticas*. pp. 13–39 Salvador: Edufba.

Renó, L., & Renó, D. (2016). Jornalismo de dados e tecnologia: algoritmo na produção da notícia transmídia | Data Journalism and Technology: Algorithm in the Production of Transmedia News. *Razón Y Palabra*, *20*(1-92), pp. 1186-1203. Retrieved from www.revistarazonypalabra.org/index.php/ryp/article/view/295

Rogers, S. (2008) "Turning Official Figures into Understandable Graphics, at the Press of a Button." *Guardian Insider Blog*. Retrieved from www.Guardian.co.uk/help/insideGuardian/2008/dec/18/unemploymentdata.

Rogers, S. (2014) Data journalism is the new punk. *British Journalism Review***,** *25(2)*, pp. 31–34. https://doi.org/10.1177/0956474814538181

Rogers, S., Gallager, A. (2013, April 4) *What is data journalism at the Guardian?* [Arquivo de Vídeo] Retrieved from www.theguardian.com/news/datablog/video/2013/apr/04/what-is-data-journalism-video.

Rothberg, D. (2008). Por uma agenda de pesquisa em democracia eletrônica. *Opinião Pública***,** *14(1)*, pp. 149–172. https://dx.doi.org/10.1590/S0104-62762008000100006.

Silva, J. A. B. e (2009) Transformações no processo de produção da notícia. *Bibliocom, 2(1)*, pp. 38–41. Retrieved from portcom.intercom.org.br/revistas/index.php/bibliocom/article/download/1524/1502

Sponholz, L. (2009). *Jornalismo, Conhecimento e Objetividade:* Além do Espelho e das Construções. Florianópolis: Insular.

DOI: 10.25200/BJR.v15n1.2019.1061

Stray, J. (2011) *A Computational Journalism Reading List*. Retrieved from jonathanstray.com/a-computational-journalism-reading-list

Träsel, M. (2014). *Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil* (PhD Dissertation) Retrieved from Repositório Institucional PUCRS (repositorio.pucrs.br/dspace/handle/10923/6841)

White House. (2009). *Open Government Directive*. Retrieved from www.digitalgov.gov/open-government-directive/

Zuiderwijk, A., Janssen, M. & Dwivedi, Y. K. (2015) Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. *Government Information Quarterly. 32 (4)*, pp. 429-440. https://doi.org/10.1016/j.giq.2015.09.005

ANA CAROLINA ARAÚJO. Substitute professor in the Department of Journalism of the Communication Faculty of the Federal University of Bahia and a PhD candidate in the Graduate Program in Communication and Contemporary Culture. Her research focuses on the areas of Public Transparency, Open Data and Identity Journalism. E-mail: contatoacaraujo@gmail.com

TRANSLATED BY: Lee Sharp

RECEIVED ON: 01/11/2017 | APPROVED ON: 16/06/2018