RESEARCH ARTICLES

Evaluating the transparency in reference data journalism. Study of the stories published between 2018 and 2019

Evaluación de la transparencia en el periodismo de datos de referencia. Estudio de las historias publicadas entre 2018 y 2019

Avaliação da transparência no jornalismo de referência. Estudo de histórias publicadas entre 2018 e 2019

Alba Córdoba-Cabús
Research staff in Training
(University of Malaga)
https://orcid.org/0000-0002-3519-0583
Spain

Manuel García-Borrego
Postdoctoral researcher
(University of Malaga)
http://orcid.org/0000-0001-6207-8741
Spain

Reception date: 25 November 2020

Review date: 6 February 2021

Accepted date: 24 May 2021

Published: 1 July 2021

To cite this article: Córdoba-Cabús, A. & García-Borrego, M. (2021). Evaluating the transparency in reference data journalism. Study of the stories published between 2018 and 2019, *Icono 14, 19*(2), 364-387. doi: 10.7195/ri14.v19i2.1648

Abstract

The approval of laws that provided access to information and promote the transparency of institutions favored the establishment of data journalism as a journalistic practice. This specialization arises in favor of transparency, so incorporating indicators that reduce opacity of the media industry is necessary to add credibility and quality to the published story. This research aims to examine the level of transparency in the best data journalism stories published worldwide between 2018 and 2019. It also aims to check whether the typology of sources and data affects opacity and to classify the pieces according to the degree of transparency. Through a descriptive and inferential analysis of the candidate projects for the Data Journalism Awards *2019 and the* Sigma Awards *2020 (n=80), a wide range of improvement in terms of transparency is detected. While the vast majority of pieces mention the source directly (91.4%), access to the methodology (48.8%) and to the data (20%) is quite limited. The multiple correspondence analysis reveals three groups of stories, many of which belong to the highest level of opacity. Among other findings, it is concluded that the type of sources and data incorporated does not influence on the level of transparency.*

Keywords: *Transparency; Data Journalism; Data; Data Journalism Awards 2019; Sigma Awards 2020; Opacity*

Resumen

La aprobación de leyes que garantizan el acceso a la información y fomentan la transparencia de las instituciones propulsó el establecimiento del periodismo de datos como práctica periodística. Esta especialización surge en favor de la transparencia, por lo que incorporar indicadores que resten opacidad a la industria mediática se plantea como necesario para añadir credibilidad y calidad a la historia publicada. Esta investigación se propone examinar el nivel de transparencia en las mejores historias de periodismo de datos reconocidas a nivel mundial publicadas entre 2018 y 2019. Asimismo, se pretende comprobar si la tipología de las fuentes y de los datos inciden en la opacidad y clasificar las piezas en función del grado de transparencia. A través de un análisis descriptivo e inferencial de los proyectos candidatos a los Data Journalism Awards 2019 *y a los* Sigma Awards 2020 *(n=80) se detecta un am-*

plio margen de mejora en cuanto a la transparencia. Mientras que en la mayoría de las piezas se menciona directamente la fuente (91,4%), el acceso a la metodología (48,8%) y a los datos en sí (20%) resulta muy limitado. Del análisis de correspondencias múltiple se desprenden tres grupos de historias, buena parte de las cuales pertenecen al nivel más opaco. Entre otras constataciones, se concluye que el tipo de fuentes y de datos incorporados no influye en el nivel de transparencia.

Palabras clave: *Transparencia; Periodismo de datos; Datos; Data Journalism Awards 2019; Sigma Awards 2020; Opacidad*

Resumo

A adopção de leis que garantem o acesso à informação e promovem a transparência das instituições favoreceu o estabelecimento do jornalismo de dados como uma prática jornalística. Esta especialização surge a favor da transparência, pelo que é necessário incorporar indicadores que reduzam a opacidade da indústria dos media para acrescentar credibilidade e qualidade à história publicada. Esta investigação visa examinar o nível de transparência nas melhores histórias de jornalismo de dados globalmente reconhecidas, publicadas entre 2018 e 2019. Visa também verificar se a tipologia das fontes e dos dados afecta a opacidade e classificar as peças de acordo com o grau de transparência. Através de uma análise descritiva e inferencial dos projectos candidatos aos Data Journalism Awards *2019 e aos* Sigma Awards *2020 (n=80), é detectada uma ampla margem de melhoria em termos de transparência. Embora a maioria das peças mencione directamente a fonte (91,4%), o acesso à metodologia (48,8%) e aos próprios dados (20%) é muito limitado. A análise de correspondência múltipla revelou três grupos de histórias, muitas das quais se encontravam ao nível mais opaco. Entre outras conclusões, conclui-se que o tipo de fontes e dados incorporados não influencia o nível de transparência.*

Palavras chave: *Transparência; Jornalismo de dados; Dados; Data Journalism Awards 2019; Sigma Awards 2020; Opacidade*

Translation by **Tridiom**

1. Introduction

Transparency is understood as a term linked to the fact that companies, governments and institutions provide users with real information of public interest, thereby fulfilling people's right to know and enabling them to make informed decisions (Lizcano, 2012). Transparency depends on the good management of organisations, which have incorporated it into their strategies with the aim of curbing the rise of mistrust among the public and the firm belief that it will have a positive impact on their reputation and the persuasive capacity of their messages (García, 2014).

In journalism, transparency is considered an essential ethical principle. Its addition, in 2014, to the code of ethics of the Society of Professional Journalists created a new paradigm for the media to face, understanding it as an opportunity to strengthen ties between journalists and institutions and the public (Kovach & Rosentiel, 2001; Deuze, 2003; Hayes, Singer & Ceppos, 2007; Karlsson, 2010; Manfredi, 2014; La-Rosa & Sandoval-Martín, 2016; Vos & Craft, 2017; Karlsson & Clerwall, 2018).

The concept of transparency in journalism, known as "media transparency" (Campos-Domínguez & Redondo-García, 2015; Díez-Garrido & Campos-Domínguez, 2018), implies revealing as much information as possible about how news is produced (Allen, 2008; Karlsson, 2010; Karlsson & Clerwall, 2018; Appelgren & Salaverría, 2018). This means that the journalist has to be honest with the audience and should explain how he/she found the information, what methods were used to analyse it and even what factors were involved in the process (Kovach & Rosentiel, 2001; Gehrke, 2020). Karlsson (2010) distinguished between two types of transparency: disclosure, which explains how the information is selected and produced, and participatory, which invites the audience to take part in the production process.

Although several academics assert that transparency affects confidence and helps to recover lost credibility (Weinberger, 2009; Singer, 2010; Philips, 2010; Vos & Craft, 2017; Anderson & Borges-Rey, 2019), others believe that its influence as a restorative element is limited (Karlsson, Clerwall & Nord, 2017; Anderson, 2018; Karlsson, 2020). Despite the discrepancies, it is plausible to think that transparency contributes effectively, especially in a situation like the present, to fighting against misleading information.

1.1. The role of transparency in data journalism

The passing of laws that guarantee access to information and promote transparency encourages the rise of data-driven projects (Cortés del Álamo, Luengo Cruz & Elías, 2018). The establishment, in part widespread, of a culture of open data led to the emergence of professional figures related to managing huge amounts of information, such as the data journalist (Paraise & Darigal, 2012; Solop & Wonders, 2016). This practice was created in favour of transparency (Beiler, Irmer & Breda, 2020) and allows us, to some extent, to move journalism away from positions that classify it as opaque (Plaisance, 2000). So much so that various authors such as Martinisi (2013), Hammond (2017), Jamil (2019) and Gehrke (2020) assume that transparency is one of the defining characteristics of data journalism, although the way in which it is implemented differs according to the media company, country and media system (Appelgren & Salaverría, 2018).

Meyer (1991) expressed the need to introduce details about the origin of information and the methodology used in data journalism. Based on these statements, transparency began to be referred to as the "new objectivity", with the understanding that this term was the evolution of traditional objectivity. Weinberger (2009) clarified that transparency was predominant in the digital age, whereas objectivity dominated the paper age. Allen (2008) later seconded Meyer's (1991) opinion and expressed the need to apply the ideals of transparency in order to boost the legitimacy of media organisations. However, as Appelgren and Salaverría (2018) suggested, for this principle to have an impact on the audience, it must be visible in the information given.

Data implies neutrality (Porter, 1995; Tal & Wansink, 2016; Kennedy, Weber & Engebretsen, 2020), but managing it requires journalists to make countless decisions. Given this control of viewpoints, providing the opportunity to reuse figures and know the methodology of the work would increase the quality of the stories and allow the watchdog role of the media to be strengthened (Philips, 2010; Alexander & Vetere, 2011; Paraise & Darigal, 2012; Howard, 2014; Coddington, 2015; Solop & Wonders, 2016; Felle, 2016; Hammond, 2017; Horky & Pelka, 2017; Anderson & Borges-Rey, 2019).

Fink and Anderson (2015) and Kennedy, Weber and Engebretsen (2020) argued that for information professionals, transparency is at the heart of data journalism. However, Solop and Wonders (2016) and Tandoc and Oh (2017) suggested that the poor recognition and appreciation of data journalism was due to the opacity of its processes. Over a decade ago, Hayes, Singer and Ceppos (2007) recognised that transparency was one of the contemporary challenges of journalism and, years later, Porlezza and Splendore (2019) maintain that it continues to be so and attribute its limited presence to journalists' own fear of having their work imitated or plagiarised.

Not enough research has been carried out on evaluating transparency in the media and those that do study it show mixed results with no consensus on established trends or patterns in newsrooms. In the last five years, there have been more studies that allude to the opacity of the news than those that reflect transparency.

Karlsson (2010), through a study of the main online versions of American, British and Swedish newspapers, showed how transparency was beginning to appear, although the degree and mode of implementation differed depending on the source of the media. On the one hand, Paraise and Darigal (2012), Tandoc y Oh (2017) and Loosen, Reimer and Schmidt (2017) found evidence of transparency in more than half of the pieces they analysed, either through the publication of raw data or incorporation of methodological details. On the other hand, Stalph (2017), Lowrey and Hou (2018), Young, Hermida and Fulda (2018), Zhang and Feng (2019) and Zamith (2019) found limited evidence of transparency in the works they studied and even classified them as opaque. Despite the difference of opinion, the studies coincide in revealing how data journalism stories are dependent on the public information shared by institutions, governments and public offices and consider this a constraint for transparency.

In short, in order to increase the reliability of journalism, we need open publications that give us an insight into the production process and data used to write the news. Given the importance of incorporating these standards into data-driven articles, this study has four objectives:

1. To determine whether there are elements of transparency in the best data journalism stories published worldwide between 2018 and 2019, taking into account the sources, data and analyses carried out.
2. To establish the properties of the transparency indicators in the projects nominated for the *Data Journalism Awards* 2019 and the *Sigma Data Journalism Awards* 2020.
3. To find out if the typology of sources and data has an impact on the level of transparency of information.
4. To classify, based on the analysis of the candidate projects for the *Data Journalism Awards* 2019 and Sigma Awards 2020, data journalism stories into groups according to their degree of transparency.

2. Materials and methods

2.1. Sample

In order to determine which pieces were considered to be of good quality in this specialist area during this period, the study focussed on analysing the candidate projects for the *Data Journalism Awards* 2019 (works published between 26 March 2018 and 7 April 2019) and the *Sigma Data Journalism Awards* 2020 (projects disseminated in 2019). Both awards, despite their different nomenclature, have a similar purpose: to celebrate the best data journalism stories around the world and honour the work of journalists. The Data Journalism Awards, organised by the Global Editors Network, took place for the last time in 2019 and was replaced by the Sigma Data Journalism Awards, founded by Aron Pilhofer and Regina Chua and sponsored by the Google News Initiative. This analysis will give us a clear idea of the current state of transparency in data journalism, without being limited by subject or geography.

For the study, duplicate works nominated in both 2019 and 2020 were identified and removed in order to avoid repetitions. Nominations for individual journalists or entire teams, applications and websites with open data were also removed from the sample. The categories chosen for analysis were: innovation, data journalism in less than 36 hours, visualisation, research and people's choice award in the 2019 competition and best data journalism story, innovation and best visualisation in the 2020 competition. The final sample was 80 projects (40 pieces from each competition).

2.2. Instrument

The way in which transparency is assessed has changed and previous studies that focussed on examining whether particular publications met this requirement do not evaluate the same categories. Hayes, Singer and Ceppos (2007) assessed transparency by analysing the sources and openness of the data, standards later replicated by Karlsson (2010), Karlsson and Clerwall (2018) and Zhang and Feng (2019) in their research. Paraise and Darigal (2012), Loosen, Reimer and Schmidt (2017), Young, Hermida and Fulda (2018) and Zamith (2019) focussed on recording access to raw data and the appearance of methodological details, while Tandoc and Oh (2017) Stalph (2017) and Lowrey and Hou (2018) continued to evaluate transparency in terms of the absence or presence of an option to download the dataset.

For this article, a descriptive and inferential quantitative content analysis was applied to the best data journalism stories published around the world between 2018 and 2019. The variables were chosen based on the observations made in scientific literature and other variables were added ad hoc in order to meet the objectives set out in this research. The codebook proposed allowed transparency to be assessed at three levels: the source of the information, the numerical characteristics and the calculations or processes carried out (Table 1).

Dimension	Variables	Categories
Sources	Number	
	Type	Public
		Private
		Other organisations
		Own sources
		Leaks
		Other
	Attribution	On the record
		On background
		Off the record
Data	Level of transparency	No source mentioned - No methodology
		Yes source mentioned - No methodology
		No source mentioned - Yes methodology
		Yes source mentioned - Yes methodology
	Access to data	No
		Yes
	Type	Geodata
		Sociodemographic
		Personal
		Measurements
		Medical
		Surveys
		Financial
		Metadata
		Other
Statistical methods	Methodology	No
		Yes
	Type of incorporation	Within the text itself
		In a specific section

Table 1: *Variables chosen for analysis. Source: own creation.*

The first dimension, sources, was evaluated based on number, type (government or public office, private companies, other organisations, leaks and own sources) and attribution (distinguishing between on the record, on background and off the record) (Hayes, Singer & Ceppos, 2007; Karlsson, 2010; Karlsson & Clerwall, 2018; Zhang & Feng, 2019).

The second level, data, evaluated the degree of transparency of the data (depending on whether the source of the information and the details of the dataset were given), if it allowed access to the raw data and the type of data (Hayes, Singer & Ceppos, 2007; Karlsson, 2010; Paraise & Darigal, 2012; Loosen, Reimer & Schmidt, 2017; Stalph, 2017; Tandoc & Oh, 2017; Lowrey & Hou, 2018; Young, Hermida & Fulda, 2018; Karlsson & Clerwall, 2018; Zamith, 2019; Zhang & Feng, 2019). At this point it was possible to discern between:

- Geodata. Specific details about geographical locations. These are provided when a map is used as a visual aid.
- Sociodemographic. Characteristics relating to a demographic group. For example: religion, income, level of education, etc.
- Personal. Specific information about an individual. For example: biography, weight, height, income, etc.
- Measurements. Values gathered using measuring tools or sensors.
- Medical. Medical information. For example: hospital discharges, admissions, ICU capacity, availability of equipment, etc.
- Surveys. Details about public opinion or points of view.
- Financial. Economic appraisals.
- Metadata. Data descriptors. For example: data about emails, use of social networks, telecommunications, etc.
- Other. Anything that does not come under the above categories.

The third level, referring to statistical processes, evaluated the incorporation of methodological information about the work with data and the way in which these were reflected in the publication (in a specific section or in the body of the text) (Paraise & Darigal, 2012; Loosen, Reimer & Schmidt, 2017; Young, Hermida & Fulda, 2018; Zamith, 2019).

This study establishes the following fundamental requirements for achieving an optimal level of transparency: directly mentioning the source or, failing that, the context from which the information originated, providing details about the characteristics of the dataset, allowing full access to the dataset and including methodological details about the work with figures.

2.3. Procedure

Based on the projects in the sample, a data matrix was created in SPSS statistical software which was used for the descriptive and inferential analyses to determine the impact of the type of sources and data used on the level of transparency of the candidate projects for the *Data Journalism Awards* 2019 and the *Sigma Data Journalism Awards* 2020.

In order to identify the link between categorical variables, the chi-square test statistic was calculated or, if this was not possible, the Yates continuity correction (if there was a degree of freedom). If the independence test was significant ($p \leq 0.05$), the adjusted residuals were examined in order to determine the direction of the link and the size of the effect caused.

In order to identify similarities between pieces of data journalism and establish groups according to transparency, a multiple correspondence analysis was chosen, which allows us to see the relationships between a set of categorical variables and depict the dimensions of our data set in a smaller space. A biplot analysis was used to explore the existing relationships between the different categories of variables and the work selected. In this case, the study was carried out using the free software R, due to the diversity and potential of the visual aids it creates.

3. Results

The findings are explained in detail below, with an emphasis on the relationships found between the study variables.

3.1. Typology of sources and attribution

The projects examined contain an average of three sources (M=2.64, SD=2.11), however, most projects incorporate information from a single source (mode=1; 29.6%). 18 projects were found in which either the exact number of sources used was not specified or a very high number were used (in some cases more than 20), so we decided to classify these as missing values and isolate them from the sample so as not to distort the measures of central tendency.

The candidate projects for the data journalism awards are based on public sources, proving a strong reliance on information from government and official offices (60%) such as the National Center for Education Statistics, Education Demographic and Geographic Estimates in the United States or the Office for National Statistics in the United Kingdom. The pieces include 36.3% of the figures sourced from organisations such as research centres, universities or NGOs, closely followed by information gathered by the media companies themselves or the bodies submitting them as candidates (35%) - this includes carrying out surveys, harvesting information from websites or documents and analysing social media. They also use, to a lesser extent, data compiled by private corporations (17.5%), figures taken from television broadcasts (included in the "other" category with 11.3%) or leaks (3.8%). One project, entitled "Key events from past four months of Hong Kong's anti-government protests"(*South China Morning Post*), was found to provide neither the exact origin of the information nor an idea of the context (1.3%).

Generally speaking, the candidates in both competitions mention the specific origin or context from which the figures were taken. Only 8.6% of the cases withhold or do not mention the identity of the source, while the source is accurately cited in the remaining cases (91.4%). Off the record attribution corresponds mostly to publications derived from leaks, such as “Driver’s notebooks exposed Argentina’s greatest corruption scandal ever: ten years and millions of cash bribes in bags” by *La Nación* (Argentina), “Copy, paste, legislate” by *Usa Today*, *The Arizona Republic* and Center for Public Integrity and “How life has changed for young people your age” by ABC News.

Type of sources	Attribution		Total	p-value
	On the record	Off the record		
No source	-	-	1.3%	-
Public Sources	95.8%	4.2%	60.0%	p=0.118
Private Sources	100%	0.0%	17.5%	p=0.969
Other organisations	96.6%	3.4%	36.3%	p=0.472
Own sources	96.4%	3.6%	35.0%	p=1.000
Leaks	0.0%	100%	3.8%	p=0.000*
Other	88.9%	11.1%	11.3%	p=1.000

Table 2: *Link between source type and attribution in nominated projects.*

News articles that use off the record attribution are based, mainly, on leaks (100%), from institutions or offices (4.2%), figures gathered by the team itself (3.6%) and other organisations (3.4%). However, the Yates continuity correction shows that the type of source used only affects attribution when it comes to leaked information (Table 2). In this sense, the relationship is close (TE=1, high effect): disseminating information of this kind requires journalists to not directly reference the source.

3.2. Type of data, transparency and access

The most frequent type of data used by the candidate projects for the data journalism awards is sociodemographic (42.5%), which is due to a large amount of data related to the characteristics of different demographic groups (income, education, employment status, etc.). Location-related data (also known as geodata) is used in a similar fashion (36.3%), closely followed by data collected by sensors or measuring tools (30%) and metadata (20%). Financial information (18.8%), personal indicators (16.3%), survey results (15%) and medical (3.8%) data are also used, although to a lesser extent.

In general, the works nominated in the 2019 and 2020 awards focus on identifying sources properly (directly or indirectly) but do not add information about the dataset (61.2%), so we cannot find out how the information is structured and what its characteristics are. This happens in pieces such as "Indonesia Plane Crash" by Reuters, “Every time Ford and Kavanaugh dodged a question, in one chart” by Vox and “Made in France” by Disclose.

Only 37.5% of the projects meet the ideal of transparency in this respect, as they include details of both the source of the figures and the database. Examples are "O que revela uma análise das emoçoes dos candidatos durante o debate” [“What an analysis of the candidates' emotions during the debate reveals”] by *O Estado de S. Paulo*, “The Invisible Crime” by *The Age*, *Sidney Morning*, *Brisbane Times* and *WA Today* and “The Quiet Rooms” by *Chicago Tribune* and Propublica Illinois. In contrast to this are the pieces that withhold the source and details of the dataset (1.3%). On this occasion only one completely opaque article was identified and that is the aforementioned "Key events from past four months of Hong Kong's anti-government protests” (*South China Morning Post*). None of the pieces were found to not state the source but include specifics on the dataset.

It is clear that the most transparent candidate projects, in that they both cite the source and provide information about the dataset, are mainly based on leaks (66.7%) and figures compiled by the team or journalist themselves (50%). This implies that, as these are pieces in which the attribution is usually off the record (leaks) and the figures are gathered or evaluated by the journalist, it is necessary to provide details about the dataset and basic information about the source in order for the publication to be classified as rigorous. Despite identifying differences in this respect, the chi-square test statistic shows that, at a confidence level of 95%, they are not significant, meaning that handling of a specific type of source or data does not influence the level of transparency achieved. However, as can be seen in Table 3, the lack of sources does imply the opacity of the article in question.

Typology	**Data transparency**[1]				**p-value**
Type of sources	0	1	2	3	
No source	100%	-	-	-	p=0.000*
Public Sources	-	64.6%	-	35.4%	p=0.395
Private Sources	-	64.3%	-	35.7%	p=0.882

Typology	**Data transparency**[1]				**p-value**
Type of sources	0	1	2	3	
Other organisations	-	69.0%	-	31.0%	p=0.468
Own sources	-	50.0%	-	50.0%	p=0.200
Leaks	-	33.3%	-	66.7%	p=0.565
Other	-	66.7%	-	33.3%	p=0.895

Table 3: *Impact of the type of source on the transparency of the information.*

The people in charge of producing publications are reluctant to make it easy to download the data from the media outlet or organisation's own website (80%). Only 20% offer this option, including “Dying homeless: counting the deaths of homeless people across the UK”, published by The Bureau of Investigative Journalism, “The Opioid Files” by the *Washington Post* and “See How the World’s Most Polluted Air Compares With Your City’s” by The *New York Times*. In these cases, the data can be downloaded in full. Projects containing data derived from leaks allow the greatest access to figures (66.7%). Meanwhile, those relying on private (85.7%) and public (83.3%) sources and including geodata (93.1%) are the most opaque. As with the previous variable, the contrast statistic shows that the incorporation of a specific type of data or source does not affect the level of data openness.

3.3. Methodological details

An examination of the presence of extra pages, pop-ups, boxes or paragraphs explaining the details of the analysis with figures shows that more than half of the articles do not include this type of information (51.2%). This is evidence of the lack of methodological details in quality data journalism pieces (lack of information on data collection, data cleaning, analysis carried out, etc.). Only 48.8% of the candidate projects for these awards include this content, two of which are "How top health websites are sharing sensitive data with advertisers" by the Financial Times and "How to profit in space: a visual guide" by The Wall Street Journal. A detailed analysis of the results shows a greater preference for incorporating this type of information in specific sections (61.4%) rather than adding it to the body of the article (18.8%).

Typology	**Methodology**^{2}			**p-value**
Type of data	0	1	2	
Geodata	51.7%	20.7%	33.3%	p=0.913
Sociodemographic	44.1%	17.6%	38.2%	p=0.375
Personal	46.2%	30.8%	23.1%	p=0.467
Medical	33.3%	-	66.7%	p=0.337
Measurements	45.8%	29.2%	25.0%	p=0.291
Encuestas	41.7%	25.0%	33.3%	p=0.741
Financieros	33.3%	6.7%	60.0%	p=0.077
Metadatos	25.0%	37.5%	37.5%	p=0.033*
Otros	66.7%	14.8%	18.5%	p=0.134

Table 4: *Impact of the type of data on the incorporation of methodology.*

A detailed methodology is less common in the candidate projects that include data related to television broadcasts (other, 66.7%), geodata (51.7%) and personal data (46.2%). When this type of data is included, it is rarely accompanied by a description of how the piece was created. At the other end of the spectrum, the publications that are most transparent in terms of incorporating a methodology are those with metadata (75% - 37.5% of which include it in the text itself and the other 37.5% include it in a specific section). With the exception of the latter, the chi-square test statistic reveals no link between the different types of data and the inclusion of a methodology (Table 4). In this sense, adding figures relating to telecommunications or about applications, emails or the use of social media (known as metadata), is intrinsically related to the appearance of methodological details in the text (TE= 0.292, moderate effect). Therefore, based on the analyses carried out in the different sections, it could be said that the type of source and data included does not affect the transparency of the information.

3.4. Types of work according to transparency

After finding that incorporating a specific type of source or data does not influence the level of transparency, a multiple correspondence analysis was carried out with the following variables: attribution, data transparency, access to data and methodology. Figure 1 shows component 1 (Dim1) has an inertia of 31.6%, while component 2 (Dim2) has an inertia of 16.6%. If we look at the X-axis, we can see a depiction of the variables in increasing order of transparency (from left to right).

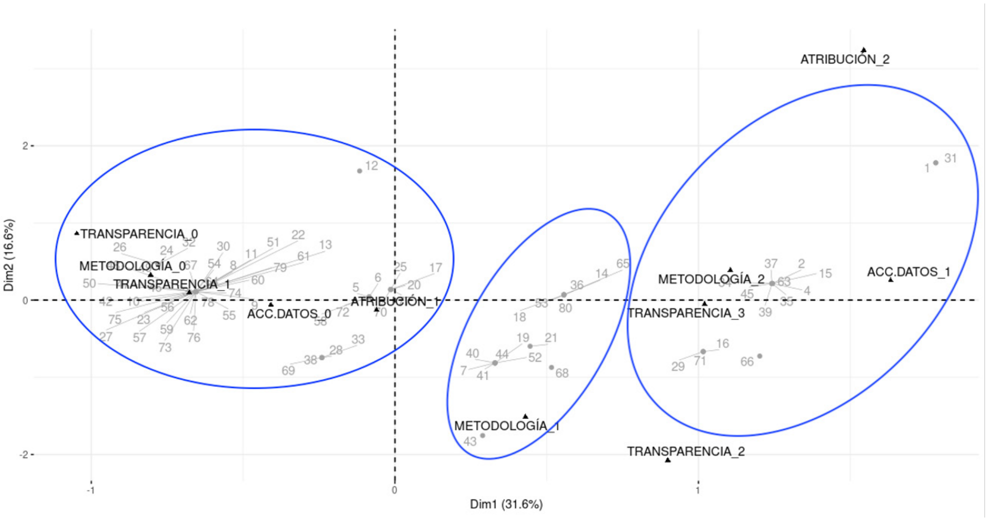

Figure 1: *Multiple Correspondence Analysis* [3].

Based on the multiple correspondence analysis, we can outline a distinction between three project groups according to their level of transparency:

- High level. This encompasses works that appear to meet the ideal of transparency described. They mention the source or, failing that, the context of the information, provide details about the dataset, specify the methodology (preferably in the body of the text) and allow access to the data.
- Medium level. This middle level contains pieces that show signs of transparency but do not comply with all the indicators. The main difference between this and the previous level is related to access to the data, as these pieces do not allow the data to be downloaded. This group includes the methodology but chooses to put it in a specific section.
- Low level. They cite the source, or come close to it, but do not specify the characteristics of the dataset or the analyses carried out and do not allow the data to be downloaded. This level also includes projects that do not mention the origin of the data. Most of the projects nominated for these awards fall into this category.

4. Discussion and conclusions

This study highlights the importance of transparency and shows that there is still plenty of room for improvement in the best data journalism stories - those nominated for the *Data Journalism Awards* 2019 and Sigma Awards 2020.

O1 and O2. Based on the results obtained, we can see that the transparency indicators are barely used, although the extent of this absence is different in each of the three viewpoints examined. While most of the works cite the sources directly, access to methodological details and to the data itself is very limited. These findings support those of authors such as Stalph (2017), Lowrey and Hou (2018), Young, Hermida and Fulda (2018) Zhang and Feng (2019) and Zamith (2019), who highlight the cautious implementation of transparency in data-driven projects. It could be argued that the dimension regarding data is the most opaque, perhaps because of a lack of resources, time and knowledge on the part of journalists to facilitate openness or, as Porlezza and Splendore (2019) point out, a fear of having their work replicated. As this study has shown, when data is available to download it can be downloaded in full and in cases that include the methodology, it appears in a specific section dedicated to it.

O3. The chi-square test statistics and the Yates continuity correction show that, in general, the typology of the sources or the data used does not affect the level of transparency of the information. However, some relationships are recorded between, for example, using data from leaks and choosing off the record attribution and handling metadata and including methodology. This would minimise the belief that insinuates that reliance on government sources is the main reason for the opacity of data journalism stories (Paraise & Darigal, 2012; Tandoc & Oh, 2017; Loosen, Reimer & Schmidt, 2017; Stalph, 2017; Lowrey & Hou, 2018; Young, Hermida & Fulda, 2018; Zhang & Feng, 2019; Zamith, 2019) and reinforce the position on the lack of resources, time and knowledge established above.

O4. The multiple correspondence factor analysis identified three groups of projects according to their level of transparency (high, medium and low). It is clear that most of the works examined correspond to the lowest level, as they prioritise mentioning sources but do not incorporate the rest of the parameters. This would prove that data journalism needs to progress in order to achieve the ideal of transparency set out in this study. Data-driven reporting should provide the transparency that journalists themselves demand from governments and institutions.

It can therefore be concluded that transparency plays a fundamental role in the production of data journalism and is partially linked to its recognition and credibility. However, what is evident from this study is that data journalism is only just starting to implement transparency and does not take advantage of the opportunities offered by websites and free resources at different stages of the production and communication process.

The type of study proposed hinders a qualitative approach to the topic in question. We aim to complement this paper in future research with in-depth interviews with data journalists in order to find out the reasons for the opacity of their work. Even so, the findings presented here reflect how works have been carried out in recent years, what characteristics they have and what the future trends will be in data journalism. These award-winning publications set future trends in this type of journalism through innovation and integrating a variety of resources. Furthermore, unlike previous studies, this study proposes a robust methodology for analysing transparency as a whole, allowing its status to be assessed and progress observed, and establishes the requirements for achieving an optimal level of transparency.

References

Alexander, S. & Vetere, C. (2011). Telling the data story the right way. *Healthcare financial Management, 65*(10), 104-110. Recovered to https://bit.ly/33dFaVe

Allen, S. D. (2008). The Trouble with Transparency. The challenge of doing journalism ethics in a surveillance society. *Journalism Studies, 9*(3), 323-40. https://doi.org/10.1080/14616700801997224

Anderson, B. & Borges-Rey, E. (2019). Encoding the UX: User Interface as a Site of Encounter between Data Journalists and Their Constructed Audiences. *Digital Journalism, 7*(9), 1253-1269. https://doi.org/10.1080/21670811.2019.1607520

Anderson, C. W. (2018). *Apostles of certainty: Data journalism and the politics of doubt*. Oxford: Oxford University Press.

Appelgren, E. & Salaverría, R. (2018). The promise of the transparency culture. A comparative study of Access to public data in Spanish and Swedish newsrooms. *Journalism Practice, 12*(8), 986-996. https://doi.org/10.1080/17512786.2018.1511823

Beiler, M., Irmer, F. & Breda, A. (2020). Data Journalism at German Newspapers and Public Broadcasters: A Quantitative Survey of Structures, Contents and Perceptions. *Journalism Studies, 21*(11), 1571-1589. https://doi.org/10.1080/1461670X.2020.1772855

Campos Domínguez, E. & Redondo García, M. (2015). Meta periodismo y transparencia informativa en el periodismo del siglo XXI. OBETS, *Revista de Ciencias Sociales, 10*(1), 185-209. https://doi.org/10.14198/OBETS2015.10.1.07

Coddington, M. (2015). Clarifying Journalism's Quantitative Turn. *Digital Journalism, 3*(3), 331-348. https://doi.org/10.1080/21670811.2014.976400

Cortés del Álamo, H., Luengo Cruz, M. & Elías, C. (2018). Periodismo de datos y transparencia al margen de los grandes medios, un estudio comparativo de Civio y Propública. *Revista ICONO14. Revista científica de Comunicación y Tecnologías emergentes, 16*(2), 66-87. https://doi.org/10.7195/ri14.v16i2.1177

Deuze, M. (2003). The Web and Its Journalism: considering the consequences of different type of news media online. *New Media & Society, 5*(2), 203-30. https://doi.org/10.1177/1461444803005002004

Díez Garrido, M. & Campos Domínguez, E. M. (2018). Los periodistas ante la transparencia en España: valoración y uso de la apertura informativa. *Revista Española de la Transparencia*, 7, 49-69. Recovered to https://bit.ly/3h0ioI7

Felle, T. (2016). Digital watchdogs? Data reporting and the news media's traditional 'fourth estate' function. *Journalism, 17*(1), 85-96. https://doi.org/10.1177/1464884915593246

Fink, K. & Anderson, C. (2015). Data Journalism in the United States. *Journalism Studies, 16*(4), 467-481. https://doi.org/10.1080/1461670X.2014.939852

García, A. M. (2014). Open government, open data, big data y transparencia: la información como nexo de unión. *COMeIN*, 39. Recovered to https://bit.ly/3kUEzOo

Gehrke, M. (2020). Transparency as a key element of data journalism. Perceptions of Brazilian professionals". En: *Computation + Journalism Symposium*. Boston: USA. Recovered to https://bit.ly/3l3jhy5

Hammond, P. (2017). From computer-computer-assisted to data-driven: Journalism and Big Data. *Journalism, 18*(4), 408-424. https://doi.org/10.1177/1464884915620205

Hayes, A., Singer, J. B & Ceppos, J. (2007). Shifting Roles, Enduring Values: the credible journalist in a digital age. *Journal of Mass Media Ethics, 22*(4), 79-262. https://doi.org/10.1080/08900520701583545

Horky, T. & Pelka, P. (2017). Data Visualization in Sports Journalism. *Digital Journalism, 5*(5), 587-606. https://doi.org/10.1080/21670811.2016.1254053

Howard, A. B. (2014). *The art and science of data-driven journalism*. New York: Tow Center for Digital Journalism. https://doi.org/10.7916/D8Q531V1

Jamil, S. (2019). Increasing Accountability Using Data Journalism: Challenges fot the Pakistani Journalists. Journalism Practice, 1-22. https://doi.org/10.1080/17512786.2019.1697956

Karlsson, M. (2010). Rituals of Transparency. *Journalism Studies, 11*(4), 535-545. https://doi.org/10.1080/14616701003638400

Karlsson, M. (2020). Dispersing the opacity of transparency in Journalism on the appeal of different forms of transparency to the general public. *Journalism Studies, 21*(13), 1795-1814. https://doi.org/10.1080/1461670X.2020.1790028

Karlsson, M. & Clerwall, C. (2018). Transparency to the Rescue? Evaluating citizens' views on transparency tools in journalism. *Journalism Studies, 19*(13), 1923-1933. https://doi.org/10.1080/1461670X.2018.1492882

Karlsson, M., Clerwall, C. & Nord, L. (2017). Do not stand corrected: transparency and users' attitudes to inaccurate news and corrections in online journalism. *Journalism & Mass Communication Quartlery, 94*(1), 148-167. https://doi.org/10.1177/1077699016654680

Kennedy, H., Weber, W. & Engebretsen, M. (2020). Data visualization and transparency in the news. En: Engebretesen, M. & Kennedy, H. (eds.). *Data Visualization in Society*. Amsterdam: Amsterdam University Press.

Kovach, B. & Rosentiel, T. (2001). *The Elements of Journalism. What news people should know and the public should expect*. New York: Crown Publishers.

La-Rosa, L. & Sandoval-Martín, T. (2016). La insuficiencia de la Ley de Transparencia para el ejercicio del Periodismo de datos en España. *Revista Latina de Comunicación Social, 71*(11), 1208-1229. https://doi.org/10.4185/rlcs-2016-1142

Lizcano, J. (2012). Transparencia. E*unomía. Revista en Cultura de la Legalidad*, 3, 160-166. Recovered to https://bit.ly/3pTVFQ6

Loosen, W., Reimer, J. & Schmidt, F. (2017). Data-driven reporting: An on-going revolution? An analysis of projects nominated for the Data Journalism Awards 2013-2016. *Journalism*, 00(0), 1-18. https://doi.org/10.1177/1464884917735691

Lowrey, W. & Hou, J. (2018). All forest, no trees? Data journalism and the construction of abstract categories. *Journalism*, 1-17. https://doi.org/10.1177/1464884918767577

Manfredi, J. L. (2014). Buenas y malas noticias sobre la ley de transparencia". *Cuadernos de periodistas*, 27, 72-80. Recovered to https://bit.ly/3fGnuH7

Martinisi, A. (2013). "Data Journalism and its role in Open Government". En: *International Scientific Conference e-Governance, Research and Educational Centre for e-Governance*. Universidad Técnica de Sofía, Sofía (Bulgaria), junio de 2013. Recovered to https://bit.ly/33dCw21 Sofía, Sofía (Bulgaria), junio de 2013. Recovered to https://bit.ly/33dCw21

Meyer, P. (1991). *The New Precision Journalism*. Indianapolis: Indiana University Press.

Paraise, S. & Darigal, E. (2012). Data-driven journalism and public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New Media & Society, 15*(6), 853-871. https://doi.org/10.1177/1461444812463345

Philips, A. (2010). Transparency and the New Ethics of Journalism. *Journalism Practice, 4*(3), 373-382. https://doi.org/10.1080/17512781003642972

Plaisance, P. L. (2000). The Concept of media Accountability Reconsidered. *Journal of Mass Media Ethics, 15*(4), 257-268. https://doi.org/10.1207/S15327728JMME1504_5

Porlezza, C. & Splendore, S. (2019). From Open Journalism to Closed Data: Data Journalism in Italy. *Digital Journalism, 7*(9), 1230-1252. https://doi.org/10.1080/21670811.2019.1657778

Porter, T.M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Singer, J. (2010). Norms and the network: Journalist ethics in a shared media space. En: Meyers, C. (Ed.). *Journalism Ethics: A philosophical approach*. (pp. 227-239). Oxford: Oxford University Press.

Solop, F. I. & Wonders, N. A. (2016). Data Journalism Versus Traditional Journalism in Election Reporting: An Analysis of Competing Narratives in the 2012 Presidential Election. *Electronic News, 10*(4), 203-223. https://doi.org/10.1177/1931243116656717

Stalph, F. (2017). Classifying Data Journalism. A content analysis of daily data-driven stories. *Journalism Practice, 12*(1), 1332-1350. https://doi.org/10.1080/17512786.2017.1386583

Tal, A. & Wansink, B. (2016). Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science, 25*(1), 117-125. https://doi.org/10.1177/0963662514549688

Tandoc, JR. & Oh, S. (2017). Small Departures, Big Continuities? Norms, values and routines in The Guardian's big data journalism. *Journalism Studies, 18*(8), 997-1015. https://doi.org/10.1080/1461670x.2015.1104260

Vos, T. P. & Craft, S. (2017). The Discursive Construction of Journalistic Transparency. *Journalism Studies, 18*(12), 1505–1522. https://doi.org/10.1080/1461670x.2015.1135754

Weinberger, D. (2009). Transparency: The new objectivity. *Knowledge Management World*. Recovered to https://bit.ly/3fw1Vc2

Young, M., Hermida, A. & Fulda, J. (2018). What Makes for Great Journalism? A content analysis of data journalism awards finalists 2012-2015. *Journalism Practice, 12*(1), 115-135. https://doi.org/10.1080/17512786.2016.1270171

Zamith, R. (2019). Transparency, Interactivity, Diversity and Information Provenance in Everyday Data Journalism. *Digital Journalism, 7*(4), 470-489. https://doi.org/10.1080/21670811.2018.1554409

Zhang, S. & Feng, J. (2019). A step forward? Exploring the diffusion of data journalism as journalistic innovations in China. *Journalism Studies, 20*(9), 1281-1300. https://doi.org/10.1080/1461670X.2018.1513814

Notes

[1] The data transparency levels are as follows: 0 = Does not mention the source or provide information about the dataset; 1 = Mentions the source but does not provide information about the dataset; 2 = Does not mention the source but provides information about the dataset; 3 = Mentions the source and provides information about the dataset.

[2] The methodology section uses the following scale: 0 = Does not include methodology; 1 = Yes, within the text itself; 2 = Yes, in a specific section.

[3] The graph shows the categories of each variable (represented by a triangle and in <variable_category> format) and the number assigned to each of the projects (represented by a dot). The variables and their corresponding categories are listed in the "Materials and methods" and "Results" sections.