Loyola University Chicago

Loyola eCommons

School of Communication: Faculty Publications and Other Works

Faculty Publications and Other Works by Department

8-25-2021

Ethical Issues in Data Journalism

Bastiaan Vanacker
Loyola University Chicago, bvanacker@luc.edu

Follow this and additional works at: https://ecommons.luc.edu/communication_facpubs

 Part of the Applied Ethics Commons, and the Journalism Studies Commons

Author Manuscript

This is a pre-publication author manuscript of the final, published article.

Recommended Citation

Vanacker, Bastiaan. Ethical Issues in Data Journalism. The Routledge Companion to Journalism Ethics, , : , 2021. Retrieved from Loyola eCommons, School of Communication: Faculty Publications and Other Works, http://dx.doi.org/10.4324/9780429262708

This Book Chapter is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in School of Communication: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.
© Bastiaan Vanacker, 2021

Ethical issues in data journalism

Bastiaan Vanacker

A process-based view on data journalism

Once the domain of computer experts and academics, analysis of large datasets has become a standard practice in newsrooms around the world. While computer-assisted reporting goes back fifty years in the United States (US), the practice of data journalism has ballooned since the late 20th century with the emergence of more affordable computers, greater availability of datasets, increased training in journalism schools, and proliferation of professional organizations such as the National Institute for Computer-Assisted Reporting (NICAR) in the US (Fink & Anderson, 2015). Elsewhere in the world, similar trends have led to increases in data journalism, as can be evidenced by the growing numbers and geographical diversity of submissions to the Data Journalism Awards (Rogers, 2019).

As boyd and Crawford (2012) observe, enormous datasets have been a reality for a long time. Big datasets, such as census data, are not novel, and some of the datasets of today's data journalism can be relatively small (such as a collection of tweets). Instead, boyd and Crawford argue that big data should be seen as a "cultural, technological, and scholarly phenomenon" (p. 663) that relies on increased computational ability to collect and analyze datasets and the belief that this will lead to a higher level of knowledge. In the

same vein, (big) data journalism should not be seen merely as journalism that uses enormous quantities of data, but as a type of journalism that relies on the increased availability of digital datasets and the ability to analyze those datasets to generate stories. This rather general definition will suffice for the purposes of this chapter.

Any data-driven journalistic project consists of four phases. In the first phase, data have to be obtained. Next, they have to be compiled and stored so that the third phase, analysis of the data, can be completed. Finally, this analysis is presented to audiences in the form of a news story. While there is some overlap between these four stages (for example, sometimes datasets can be obtained from organizations that have already compiled and stored the data), they provide a useful framework to lay out the various ethical issues that can arise during a data-driven journalistic project.

Phase one: Collection

As with all journalistic projects, data journalism starts with getting the facts. This can entail painstakingly collecting information and plugging it into spreadsheets, or it can be as easy as obtaining a database compiled by a government agency. Issues related to accuracy, source reliability, independence, and ethical newsgathering methods are most likely to come up in this phase.

Checking for accuracy and reliability

As mentioned above, one of the driving forces behind data journalism has been the increasing availability of data. The "data portals" many cities in the US have launched during the last decade have been an important source of data for local reporters (Thornton, 2013). These portals contain all types of government data that range from pothole repairs to salaries of city workers. But journalists tend to be skeptical of information produced by the government and often will do additional reporting to verify it; should they also do this with the kind of information provided via government-maintained data portals? According to Chicago data journalist Elliott Ramos (2013), some of his colleagues do indeed distrust the data from the Chicago Data Portal. Data that have important political consequences, such as crime incidents or police brutality cases, need to be approached particularly critically. Parasie and Dagiral (2013) note that editors at the *Chicago Tribune* decided not to use crime data provided by the Chicago Police Department in compiling a homicide map and instead did additional reporting to obtain more accurate information.

In many cases, however, checking the accuracy of all the data in a dataset can be practically or logistically impossible. One solution hinted at by Ramos (2013) is to treat this data as a quote from a government agency in consecutive reporting, so that the news outlet is not directly accountable for false information. It is also worth noting that not all government data are provided through these portals, and portals do not eliminate the need for other ways to obtain data, such as – in the US – filing a request under a state's open record laws.

Should reporters use hacked data?

When hackers make illegally obtained datasets publicly available, should reporters feel free to use them? The journalistic approach to this issue has been that once these data are dumped, there is no objection to accessing and using them; at this point, the proverbial horse has left the barn. Consider these two examples: In 2014, mainstream media considered hacked and released emails from Sony executives in which they made snide remarks about movie stars to be fair game. Additionally, the media widely reported on the emails of Hillary Clinton's 2016 campaign chair John Podesta that were released on Wikileaks. The information contained in both of these dumps was newsworthy and accurate, and most media organizations felt confident using them.

But another view is that these datasets are tainted by the ways in which they were obtained. The Podesta emails, for example, have been alleged to be hacked by Russian intelligence agents. The Sony hack is thought to have been the work of North Korean hackers. Some have argued that by using the information provided through these data breaches, journalists are complicit in the hack. Aaron Sorkin (2014) famously criticized the media for publishing emails in which Sony executives are squabbling and spreading industry gossip:

> If you close your eyes you can imagine the hackers sitting in a room, combing through the documents to find the ones that will draw the most blood. And in a room next door are American journalists doing the same thing. As demented and

> criminal as it is, at least the hackers are doing it for a cause. The press is doing it for a nickel (para. 17).

Similarly, Zittrain (2016) argues that news organizations need to cease treating information dumps that contain email conversations of public officials the same way they approach information fed to them by a source or a whistle-blower. These hacks are selective and meant to intervene in our public discourse in a way that suits the hackers' needs, he argues.

This raises the question: Should the source's motivation enter the moral decision-making process? In the (American) context of anonymous sources, Duffy and Freeman (2011) observe that media ethics textbooks tend to ignore sources' motives. Similarly, Walter Pincus (2005), who reported until 2015 on national security issues for *The Washington Post*, argues that only three considerations come in to play when contemplating printing classified information gathered from an anonymous source: Is it newsworthy? Is it accurate? Does reporting on this harm national security? To journalists, it seems, information is information, and the provenance of information is relevant in assessing its validity and accuracy, but the motivations of sources do not enter the moral equation. In an era when government-funded or ideologically motivated hacker collectives selectively expose data in an attempt to influence policy or elections, this ethical approach might need revisiting.

To scrape or not to scrape

In the examples above, I discussed the use of third-party databases provided to journalists, but reporters also create their own datasets. “Scraping,” or using software to save large swaths of data from websites, is one way in which they can do this. Many site owners dislike the practice because they consider the information as proprietary or because they simply prefer to interact with “real” visitors instead of bots. But does this mean that scraping as a practice is unethical? Mowle (2019) argues that if people subscribe to a service that bars scraping, they might understand that their data will not be subjected to it. Interestingly, Mowle herself, when scraping historical Twitter data in the context of her academic research, did not think she was bound by Twitter’s terms of service since she herself was not a Twitter user and, therefore, had not subscribed to these terms. However, if reporters consent to a terms-of-service agreement that bars scraping in order to get access to a site or use a service, they should abide by it and not try to get access to data that are not freely available (Shiab, 2018).

Scraping software can slow a site down and can be designed to disguise itself as a regular visitor to the site owner. Densmore (2017) articulates a number of rules for ethical scraping that require that the software does not impact the site’s performance and that the data will only be harvested through scraping if no other option is available. For example, sometimes sites make their data and functionality open to other developers through an Application Programming Interface (API). Densmore’s rules also require that scraping software allows site owners to figure out what it is going on and provide them with the contact information of the person collecting the data. Shiab (2018) compares scraping

without identifying oneself to undercover reporting, a practice that is considered to be an exception to standard journalistic practice.

Researchers or journalists scrape web content by developing a software program and instructing that program to copy and store a web site's content. A coding error in the scraping software can lead to mistakes in analyzing and reporting the harvested data. For example, if one were to analyze the average selling price of homes in a zip code by scraping online real estate transaction records, the program could generate unreliable results if it erroneously scraped the amount of the mortgage taken out instead of the actual amount a house sold for. Or if it accidentally arranged houses by zip code of the buyer instead of the seller, results would be skewed. Making the code publicly available allows others to verify and assess the quality of data gathered and engage in some sort of review process. The ethos of collaboration that pervades the open source community also mandates sharing scripts in the spirit of collaboration. However, some journalists are reluctant to do so, especially if they put time and effort in developing the script (Shiab, 2018).

Paying for datasets

Bradshaw (2015) questions whether journalists should pay for source material such as datasets. In the US, there is a long-standing tradition that journalists do not pay their sources. Should this preclude reporters from paying for data? According to the Society of Professional Journalists, there are a number of reasons why journalists should not engage in this practice (Farrell, n.d.). The reasons for this boil down to the fact that checkbook

journalism mars and corrupts the source-reporter relationship in a way that brings long-term harm to journalism. These criticisms also assume that checkbook journalism is associated with exclusivity, allowing reporters to possess information their competitors lack. These reasons seem valid if a source were to peddle a dataset to a journalist for money, but in the case of commercially produced datasets that are for sale to everyone, these arguments do not seem to hold. In that instance, buying databases might be more akin to getting a subscription to Lexis-Nexis than paying off a source.

Phase two: Compilation

In many cases, data can be obtained as datasets that can then be analyzed. But this is not always the case, and in many instances, data reporters have to enter or re-enter the data in a database before they can be processed.

Adhering to fair information practices

In instances where a media organization creates a database that contains personally identifiable information, legal and ethical requirements regarding privacy and security should be observed. Particularly when dealing with sensitive data, media organizations need to minimize security risks and ensure secure storage. Around the world, most notably in Europe, data protection laws have become stricter. While these laws tend to have exceptions for journalism, the fair information practices underpinning these laws (which were developed in the US by the Department of Health, Education, and Welfare)

can serve as ethical guideposts for media organizations creating and maintaining databases.

These principles state, among other things, that consent is required for the collection of the data, and that this consent only applies to the use of the data for the specific purpose for which they were collected. For example, people might consent to having their data stored by Facebook in order to enjoy the service, but they might not consent to having their personally identifiable information used by political campaigns. Aggregating and anonymizing the data addresses only some of these concerns. The mere collection of data, aside from their use, presents a privacy issue. Especially when dealing with financial information and other sensitive information, data security is of the utmost importance. For example, while investigating and sharing with one another the leaked financial documents at issue in the Panama Papers investigations, journalists from around the world successfully adhered to strict security protocols (Romera & Gallego, 2018) that Gillian Phillips from the Guardian and Observer Media Group describes in this volume.

Phase three: Analysis

Once data have been collected and stored, it is up to the data journalist to discover the stories hidden within these data. In doing so, reporters have to ensure that their analyses comply with journalistic standards of objectivity and accuracy. Numbers create an illusion of precision and accuracy than can conceal hidden biases and methodological errors.

Accounting for errors

Above, I discussed concerns regarding intentional manipulation of datasets. However, there are many other less conspicuous ways that datasets can be "dirty," i.e. contain incomplete or erroneous data. For example, when data are missing, a default value might be entered, skewing the results. A trained data reporter, however, will pick up on this and spot the outliers in the respective datasets. By dissecting a much-shared online story about pornography consumption and political affiliation, Harris (2014) illustrates the need for data literacy. The story indicated that, on a per capita basis, states carried by Barrack Obama in the 2012 election consumed more pornography than states that voted for Mitt Romney. This analysis was based on data provided by Pornhub, which had geocoded IP addresses from its traffic log. However, this analysis was based on a number of assumptions, for example that Pornhub page views are a valid way to measure porn consumption in general and that geocoding is reliable. And as Harris points out, these assumptions are flawed. Many IP addresses only allow geocoding at the country, not the state level. When this happened, the program assigned Kansas as the location of the IP address, creating the impression that people from Kansas have a voracious appetite for pornography. The analysis also ignored the fact that other intervening variables – for example, faster internet speeds in urban areas – might account for the difference, mistaking correlation for causation.

Data are the result of a process in which facts and information are coded, so they can be manipulated by software. This process is prone to error at various points in the chain.

Errors can stem from inconsistent coding, problems with data import, misleading default settings, flawed information, and a host of other problems too numerous to include in this overview. Seasoned data reporters don't ask if their data are dirty, but how dirty they are, and then they try to clean that data up. Nevertheless, for many journalists, the problem with dirty data remains under the radar (Messner & Garrison, 2007).

Acknowledging bias

Through the work of scholars such as O'Neil (2016), we have learned about the biases baked into databases and the technology that we use to interpret them. Everything can be quantified, but this process of quantification is not neutral. When using data, reporters should, therefore, question whether they are perpetuating biases rather than exposing them. For example, it is a well-established fact that police arrest black people at much higher rates than people of other races in the US (Schleiden, Soloski, Milstead & Rhynehart, 2020). News organizations that take arrest records and mugshots at face value as a basis for analysis enforce these biases and are complicit in perpetuating structural inequalities.

Phase four: Use and presentation

After the data have been collected and analyzed, they are shaped into a story, requiring editors to make ethical decisions about how the data (analysis) should be shared with audiences. This requires striking a balance between providing transparency, respecting privacy, and protecting proprietary information.

Considering amplification

In the US, salaries of public employees, how much people paid for their homes (and how much they loaned), or how much they pay in property taxes can often be easily looked up or requested. But if reporters amplify this information by lifting it from rarely used databases or paper records into their media outlets without a clear journalistic purpose, they open themselves up to criticism.

For example, two weeks after the shooting at Sandy Hook Elementary School in 2012, the *Journal News* in White Plains, New York, published three clickable maps of gun permit owners in two counties in suburban New York for a story on access to gun permit data (Craig, Ketterer & Yousuf, 2017). The maps contained the names and addresses of gun owners as well as the permit information. This decision caused an outcry, as many argued that there was no journalistic purpose that justified invading the privacy of gun owners and exposing them to potential harms (such as making them targets for burglaries). Bartzen Culver (2013) argues that the newspaper did not account for the fact that there might have been inaccuracies in the database and that without proper context, this dataset failed to provide journalism of public interest. As such, the data did not tell a story about crime, gun ownership, or flaws in the gun permit system. As media ethicist Al Tompkins (2012) notes: “If publishing the data because it is public and the public seems to be interested in the topic right now is reason enough, then there are endless databases to exploit ” (para. 21).

When David Dao's forceful removal from an overbooked United Airlines flight in 2017 was captured by his co-passengers, their recordings of the event made national headlines. While most of the coverage focused on airlines' overbooking procedures, his hometown paper, *The Louisville Courier-Journal*, relied on public documents to reveal some troubling details from his past, including a series of drug-related offenses (Watkins, 2017). Many critics found this irrelevant to the story and a poor use of public records (Lakshmanan, McBride & Tompkins, 2017). A television reporter who posted a picture on Twitter of all the court and legal documents she had amassed on the doctor was vilified, and she quickly removed the tweet (Concha, 2017).

These examples show that publicly accessible data are not necessarily newsworthy or fair game. Just as not every detail from a police report on a murder would make it into a news report, neither should every detail of a document search. Before publishing personal data of this nature, journalists should ask whether doing so fulfills a journalistic purpose, consider the harm of posting the information, and think about alternatives (Craig, Ketterer & Yousuf, 2017).

Providing transparency

As discussed above, databases are prone to error. And despite journalists' best efforts to clean up data, some mistakes might slip through the cracks. Should media organizations, therefore, share their datasets with audiences, so those audiences can check for themselves? Within the civic tech movement, where sharing and collaborating are core values, as well as in the social sciences with their peer review requirements, this is

standard practice. By sharing their data, news organizations can enable and encourage others to find new story angles, visualizations, or applications for the data. However, news organizations do not always look at it this way. They might have spent considerable resources on filing public record requests from reluctant government agencies or on scraping web sites. Sharing this information so others can take advantage of the news organizations' hard work might not be an appealing option. Moreover, if databases contain personal information, it would require ensuring that this information is redacted or removed, which, of course, also requires time and resources.

Some news organizations such as ProPublica make their datasets available for costs ranging from a few hundred to a few thousand dollars. ProPublica's senior editor for news applications said this about the practice: "If another newsroom does great journalism using our material, it's a great day in the office…In a traditional model that would be called 'getting scooped by our own story,' but that's a huge success for us" (Ali, 2014, para. 11). This attitude indicates that the ethos of the open source movement is permeating journalism, even though, at this point in time, many traditional news organizations might feel less inclined to share their data with their competitors.

Conclusion

This overview of the ethical issues facing data-driven journalism in each phase of the reporting process is necessarily limited and incomplete. First, the confines of a book chapter do not allow to cover the myriad potential ethical issues that can emerge in the

context of a data-driven journalistic project. These confines forced me to make a selection, and every selection necessarily implies elimination. Second, this overview is incomplete because the ethical issues, along with breakthroughs in technology, are rapidly evolving. For example, this chapter limited itself to data in the traditional sense – i.e. information contained in documents and spreadsheets – but it ignored data that can be generated by facial recognition technology or predictive analytics, to name but two.

As the technology keeps evolving, so will the ethical issues. However, a discerning reader might have observed that despite the seemingly novel nature of some of the issues observed in this chapter, they also are very traditional. At their core, these issues center around familiar journalistic values such as telling the truth while minimizing harm and being accountable while remaining independent. And while evolving technologies and reporting techniques will continue to bring up new ethical dilemmas, these well-established values will remain a trusted resource for reporters looking for a moral anchor.

Further reading

Bradshaw, Paul. (2015). Data journalism. In L. Zion & D. Craig (Eds.), *Ethics for digital journalists: emerging best practices* (pp. 200–219). New York; London: Routledge, Taylor & Francis Group.

Gray, J., Chambers, L., & Bounegru, L. (Eds.) (2012). *The data journalism handbook: How journalists can use data to improve the news*. Sebastopol, CA: O'Reilly Media, Inc.

Harris, J. (2014, May 22). Distrust your data. *Source*. https://source.opennews.org/articles/distrust-your-data/

References

Ali, T. (2014, April 28). ProPublica plans to grow its "Data Store." *Columbia Journalism Review*. http://www.cjr.org/behind_the_news/propublica_plans_to_grow_its_d.php

Bartzen Culver, K. (2013, February 22). Where the Journal News went wrong in mapping gun owners. *Mediashift.* http://mediashift.org/2013/02/where-the-journal-news-went-wrong-in-mapping-gun-owners053/

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Bradshaw, Paul. (2015). Data journalism. In L. Zion & D. Craig (Eds.), *Ethics for digital journalists: emerging best practices* (pp. 200–219). New York; London: Routledge, Taylor & Francis Group.

Concha, J. (2017, April 12). Why did the media try to smear United's beating victim? *The Hill*. https://thehill.com/blogs/pundits-blog/media/328410-why-did-the-media-try-to-smear-uniteds-beating-victim

Conversation with data #26: Ethical dilemmas in data journalism. (2019, May). *Data Journalism*. https://datajournalism.com/read/newsletters/ethical-dilemmas-in-data-journalism

Craig, D., Ketterer, S., & Yousuf, M. (2017). To post or not to post: Online discussion of gun permit mapping and the development of ethical standards in data journalism. *Journalism & Mass Communication Quarterly*, *94*(1), 168–188.

Densmore, J. (2017, July 23). Ethics in web scraping. https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

Duffy, M. J. & Freeman, C. P. (2011). Anonymous sources: A utilitarian exploration of their justification and guidelines for limited use. *Journal of Mass Media Ethics, 26*(4), 297-315.

Farrell, M. (n.d.). Checkbook journalism. *Society of Professional Journalists.* https://www.spj.org/ethics-papers-cbj.asp

Fink, K., & Anderson, C. W. (2015). Data Journalism in the United States: Beyond the "usual suspects." *Journalism Studies*, *16*(4), 467–481.

Harris, J. (2014, May 22). Distrust your data. *Source*. https://source.opennews.org/articles/distrust-your-data/

Lakshmanan, I., Tompkins, A., & McBride, K. (2017, April 11). Was the United passenger's 'troubled past' newsworthy? *Poynter*. https://www.poynter.org/ethics-trust/2017/was-the-united-passengers-troubled-past-newsworthy/

Messner, M., & Garrison, B. (2007). Journalism's 'Dirty Data' below researchers' radar. *Newspaper Research Journal*, *28*(4), 88–100.

Mowle, A. (2019, May). Conversations with data: #26. Ethical dilemmas in data journalism. *Datajournalism.com.* https://datajournalism.com/read/newsletters/ethical-dilemmas-in-data-journalism

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition). New York: Crown Books.

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), 853–871.

Pincus, W. (2005, June 15) Anonymous sources: Their use in a time of prosecutorial interest. *Nieman Reports*. https://niemanreports.org/articles/anonymous-sources-their-use-in-a-time-of-prosecutorial-interest/

ProPublica Data Store. (2019, June 10). https://www.propublica.org/datastore

Ramos, E. (2013). A journalist's take on open data. In B. Goldstein & L. Dyson (Eds.), *Beyond transparency: open data and the future of civic innovation* (pp. 93–104). https://beyondtransparency.org/pdf/BeyondTransparency.pdf

Rogers, S. Data journalism becomes a global field. *Niemanlab Predictions for 2019.* https://www.niemanlab.org/2018/12/data-journalism-becomes-a-global-field/

Romera, P., & Galego, C. S. (2018, July 3). How ICIJ deals with massive data leaks like the Panama Papers and Paradise Papers. *Consortium of Investigative Journalists* https://www.icij.org/blog/2018/07/how-icij-deals-with-massive-data-leaks-like-the-panama-papers-and-paradise-papers/

Schleiden, C., Soloski, K. L., Milstead, K., & Rhynehart, A. (2020). Racial disparities in arrests: a race specific model explaining arrest rates across black and white young adults. *Child and Adolescent Social Work Journal*, *37*(1), 1-14.

Shiab, N. (2015, August 12). On the ethics of web scraping and data journalism. *Global Investigative Journalism Network*. https://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/

Sorkin, A. (2014, December 14). The Sony hack and the yellow press. *The New York Times*. https://www.nytimes.com/2014/12/15/opinion/aaron-sorkin-journalists-shouldnt-help-the-sony-hackers.html

Thornton, S. (2013, September 30). Open data in Chicago. *Data-Smart City Solutions*. https://datasmart.ash.harvard.edu/news/article/open-data-in-chicago-a-comprehensive-history-311

Tompkins, A. (2012, December 27). Where the Journal News went wrong in publishing names, addresses of gun owners. *Poynter*. https://www.poynter.org/reporting-editing/2012/where-the-journal-news-went-wrong-in-publishing-names-addresses-of-gun-owners/

Watkins, M. (2017, April 11). David Dao, passenger removed from United flight, now in spotlight. *The Courier-Journal*. https://www.courier-journal.com/story/news/local/2017/04/11/david-dao-passenger-removed-united-flight-doctor-troubled-past/100318320/

Zittrain, J. (2016, October 19). Mass hacks of private email aren't whistleblowing, they are at odds with it. *Just Security*. https://www.justsecurity.org/33677/mass-hacks-private-email-arent-whistleblowing-odds-it/