20. Working Openly in Data Journalism

Natalia Mazotte

Abstract

This chapter examines some examples and benefits of data journalists working more openly, as well as some ways to get started.

Keywords: data journalism, open source, free software, transparency, trust, programming

Many prominent software and web projects—such as Linux, Android, Wikipedia, WordPress and TensorFlow—have been developed collaboratively based on a free flow of knowledge.¹ Stallman (2002), a noted hacker who founded the GNU Project and the Free Software Foundation, says that when he started working at MIT in 1971, sharing software source code was as common as exchanging recipes.

For years such an open approach was unthinkable in journalism. Early in my career as a journalist, I worked with open-source communities in Brazil and began to see openness as the only viable path for journalism. But transparency hasn't been a priority or core value for journalists and media organizations. For much of its modern history, journalism has been undertaken in a paradigm of competition over scarce information.

When access to information is the privilege of a few and when an important finding is only available to eyewitnesses or insiders, ways of ensuring accountability are limited. Citing a document or mentioning an interview source may not require such elaborate transparency mechanisms. In some cases, preserving the secrecy means ensuring the security of the source, and is even desirable. But when information is abundant, not sharing the *how-we-got-there* may deprive the reader of the means to understand and make sense of a story.

1 This chapter was written by Natalia Mazotte with contributions from Marco Túlio Pires.

As journalists both report on and rely on data and algorithms, might they adopt an ethos which is similar to that of open-source communities? What are the advantages of journalists who adopt emerging digital practices and values associated with these communities? This chapter examines some examples and benefits of data journalists working more openly, as well as some ways to get started.²

Examples and Benefits of Openness

The Washington Post provided an unprecedented look at the prescription opioid epidemic in the United States by digging into a database on the sales of millions of painkillers.³ They also made the data set and its methodology publicly accessible. This enabled local reporters from over 30 states to publish more than 90 articles about the impact of this crisis in their communities (Sánchez Díez, 2019).⁴

Two computational journalists analyzed Uber's surge pricing algorithm and revealed that the company seems to offer better service in areas with more White people (Stark & Diakopoulos, 2016). The story was published by *The Washington Post*, and the data collection and analysis code used were made freely available on GitHub, an online platform that helps developers store and manage their code.⁵ This meant that a reader who was looking at the database and encountered an error was able to report this to the authors of the article, who were in turn able to fix the bug and correct the story.

Gênero e Número (Gender and number), a Brazilian digital magazine I co-founded, ran a project to classify more than 800,000 street names to understand the lack of female representation in Brazilian public spaces. We did this by running a Python script to cross-reference street names with a database of names from the Brazilian National Statistical office (Mazotte & Justen, 2017). The same script was subsequently used by other initiatives to classify data sets that did not contain gender information—such as lists of electoral candidates and magistrates (Justen, 2019).

Working openly and making various data sets, tools, code, methods and processes transparent and available can potentially help data journalists in a

2 For more on data journalism and open-source, see also chapters by Leon, Baack, and Pitts and Muscato in this book.

3 <https://www.washingtonpost.com/national/2019/07/20/opioid-files/>

4 <https://www.washingtonpost.com/national/2019/08/12/post-released-deas-data-pain-pills-heres-what-local-journalists-are-using-it/>

5 <https://github.com/comp-journalism/2016-03-wapo-uber>

number of ways. Firstly, it can help them to improve the quality of their work. Documenting processes can encourage journalists to be more organized, more accurate and less likely to miss errors. It can also lighten the burden of editing and reviewing complex stories, enabling readers to report issues. Secondly, it can broaden reach and impact. A story that can be built upon can gain different perspectives and serve different communities. Projects can take on a life of their own, no longer limited by the initial scope and constraints of their creators. And thirdly, it can foster data literacy amongst journalists and broader publics. Step-by-step accounts of your work mean that others can follow and learn—which can enrich and diversify data ecosystems, practices and communities.

In the so-called “post-truth” era there is also potential to increase public trust in the media, which has reached a new low according to the 2018 Edelman Trust Barometer.⁶ Working openly could help decelerate or even reverse this trend. This can include journalists talking more openly about how they reach their conclusions and providing more detailed “how tos,” in order to be honest about their biases and uncertainties, as well as to enable conversations with their audiences.⁷

As a caveat, practices and cultures of working openly and transparently in data journalism are an ongoing process of exploration and experimentation. Even as we advance our understanding of potential benefits, consideration is needed to understand when transparency is valuable, or might be less of a priority, or might even be harmful. For example, sometimes it’s important to keep data and techniques confidential in order to protect the integrity of the investigation itself, as happened in the case of the Panama Papers.

Ways of Working Openly

If there are no impediments (and this should be analyzed on a case-by-case basis) then one common approach to transparency is through the methodology section, also known as the “nerd box.” This can come in a variety of formats and lengths, depending on the complexity of the process and the intended audience.

6 <https://www.edelman.com/research/2018-edelman-trust-barometer>; https://www.edelman.com/sites/g/files/aatuss191/files/2018-10/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf

7 For more on issues around uncertainty in data journalism, see Anderson’s chapter in this volume.

If your intention is to reach a wider audience, a box inside the article or even a footnote with a succinct explanation of your methods may be sufficient. Some publications opt to publish stand-alone articles explaining how they reported the story. In either case, it is important to avoid jargon, explain how data was obtained and used, ensure readers don't miss important caveats, and explain in the most clear and direct way how you reached your conclusion.

Many media outlets renowned for their work on data journalism—such as *FiveThirtyEight*, *ProPublica*, *The New York Times* and the *Los Angeles Times*—have repositories on code-sharing platforms such as GitHub. The *Buzzfeed News* team even has an index of all its open-source data, analysis, libraries, tools and guides.⁸ They release not only the methodology behind their reporting, but also the scripts used to extract, clean, analyze and present data. This practice makes their work reproducible (as discussed further in Leon's chapter in this volume) as well as enabling interested readers to explore the data for themselves. As scientists have done for centuries, these journalists are inviting their peers to check their work and see if they can arrive at the same conclusions by following the documented steps.

It is not simple for many newsrooms to incorporate these levels of documentation and collaboration into their work. In the face of dwindling resources and shrinking teams, journalists who are keen to document their investigations can be discouraged by their organizations. This brings us to the constraints that journalists face: Many news organizations are fighting for their lives, as their role in the world and their business models are changing. In spite of these challenges, embracing some of the practices of free and open-source communities can be a way to stand out, as a marker of innovation and as a way of building trust and relationships with audiences in an increasingly complex and fast-changing world.

Works Cited

- Justen, A. (2019, May 31). Classificando nomes por gênero usando dados públicos | Brasil.IO—Blog. *Brasil.IO*. <https://blog.brasil.io/2019/05/31/classificando-nomes-por-genero-usando-dados-publicos/>
- Mazotte, N., & Justen, A. (2017, April 5). Como classificamos mais de 800 mil logradouros brasileiros por gênero. *Gênero e Número*. <http://www.generonumero.media/como-classificamos-mais-de-800-mil-logradouros-brasileiros-por-genero/>

8 <https://github.com/BuzzFeedNews/everything#guides>

- Sánchez Díez, M. (2019, November 26). *The Post* released the DEA's data on pain pills. Here's what local journalists are using it for. *The Washington Post*. <https://www.washingtonpost.com/national/2019/08/12/post-released-deas-data-pain-pills-heres-what-local-journalists-are-using-it/>
- Stallman, R. M. (2002). *Free software, free society: Selected essays of Richard M. Stallman* (J. Gay, Ed.). GNU Press.
- Stark, J., & Diakopoulos, N. (2016, March 10). Uber seems to offer better service in areas with more White people. That raises some tough questions. *The Washington Post*. <https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/>

About the Author

Natalia Mazotte is a Brazilian data journalist and entrepreneur. She co-founded the news start-up *Gênero e Número* and the School of Data Brazil.