38. From *The Guardian* to Google News Lab: A Decade of Working in Data Journalism

Simon Rogers

Abstract

A personal narrative of the last decade of data journalism through the lens of the professional journey of one of its acclaimed figures.

Keywords: data journalism, *The Guardian*'s Datablog, WikiLeaks, open data, transparency, spreadsheets

When I decided I wanted to be a journalist, somewhere between the first and second years of primary school, it never occurred to me that would involve data. Now, working with data every day, I realize how lucky I was. It certainly was not the result of carefully calibrated career plans. I was just in the right place at the right time. The *way* it happened says a lot about the state of data journalism in 2009. I believe it also tells us a lot about data journalism in 2019.

Adrian Holovaty, a developer from Chicago who had worked at *The Washington Post* and started EveryBlock, a neighbourhood-based news and discussion site, came to give a talk to the newsroom in the Education Centre of *The Guardian* on Farringdon Road in London. At that time I was a news editor at the print paper (then the centre of gravity), having worked online and edited a science section. The more Holovaty spoke about using data to both tell stories and help people understand the world, the more something triggered in me. Not only could I be doing this, but it actually reflected what I *was* doing more and more. Maybe I could be a journalist who worked with data. A “data journalist.”

Working as a news editor with the graphics desk gave me the opportunity to work with designers who changed how I see the world, in Michael

Robinson's talented team. And as the portfolio of visuals grew, it turned out that I had accumulated a lot of numbers: Matt McAlister, who was launching *The Guardian's* open API, described it as "the motherlode." We had GDP data, carbon emissions, government spending data and much more cleaned up, all saved as Google spreadsheets and ready for use the next time we needed it.

What if we just published this data in an open data format? No PDFs, just interesting accessible data, ready to use, by anyone. And that's what we did with *The Guardian's* Datablog—at first with 200 distinct data sets: Crime rates, economic indicators, war zone details, and even fashion week and Doctor Who villains. We started to realize that data could be applied to everything. It was still a weird thing to be doing. "Data editor" was hardly a widespread job—very few newsrooms had any kind of data team at all. In fact, just using the word "data" in a news meeting would elicit sniggers. This wasn't "proper" journalism, right?

But 2009 was the start of the open data revolution: US government data hub data.gov had been launched in May of that year with just 47 data sets. Open data portals were being launched by countries and cities all over the world, and campaigners were demanding access to ever more. Within a year, we had our readers helping to crowdsource the expenses of thousands of MPs. Within the same period, the UK government had released its ultimate spending data set: COINS (Combined Online Information System) and *The Guardian* team had built an interactive explorer to encourage readers to help explore it.¹ Once stories were produced from that data, however, the ask became, "How can we get more of this?"

There wasn't long to wait. The answer came from a then-new organization based in Sweden with what could charitably be described as a radical transparency agenda: WikiLeaks. Whatever you feel about WikiLeaks today, the impact of the organization on the recent history of data journalism cannot be overstated. Here was a massive dump of thousands of detailed records from the war zones of Afghanistan first, followed by Iraq. It came in the form of a giant spreadsheet, one too big for the investigations team at *The Guardian* to handle initially.

It was larger than the Pentagon Papers, that release of files during the Vietnam War which shed light on how the conflict was really going. The records were detailed too—including a list of incidents with casualty counts, geo locations, details and categories. We could see the rise in IED attacks in Iraq, for instance, and how perilous the roads around the country had

1 <https://www.theguardian.com/politics/coins-combined-online-information-system>

become. And when that data was combined with the traditional reporting skills of seasoned war reporters, the data changed how the world saw the wars.

It wasn't hard to produce content that seemed to have an impact across the whole world. The geodata in the spreadsheets lent itself to mapping, for instance, and there was a new free tool which could help with that: Google Fusion Tables. So we produced a quick map of every incident in Iraq in which there had been at least one death. Within 24 hours, a piece of content which took an hour to make was being seen around the world as users could explore the war zone for themselves in a way which made it seem more real. And because the data was structured, graphics teams could produce sophisticated, rich visuals which provided more in-depth reporting.

And by the end of 2011—the year before this book was first published—the “Reading the Riots” project had applied the computer-assisted reporting techniques of Phil Meyer in the 1960s to an outbreak of violence across England (Robertson, 2011). Meyer had applied social science techniques to reporting on the Detroit riots of the late 1960s. A team led by *The Guardian's* Paul Lewis did the same to the outbreak of unrest across England that year and incorporated data as a key part of that work. These were front-page, data-based stories.

But there was another change happening to the way we consume information, and it was developing fast. I can't remember hearing the word “viral” outside health stories before 2010. The same is not true today and the rise of data journalism also coincided with the rise of social media. We were using tweets to sell stories to users across the globe and the resultant traffic led to more users looking for these kinds of data-led stories. A visual or a number could be seen in seconds by thousands. Social media transformed journalism but the amplification of data journalism was the shift which propelled it from niche to mainstream.

For one thing, it changed the dynamic with consumers. In the past, the words of a reporter were considered sacrosanct; now you are just one voice among millions. Make a mistake with a data set and 500 people would be ready to let you know. I can recall having long (and deep) conversations on Twitter with designers around colour schemes for maps—and changing what I did because of it. Sharing made my work better.

In fact that spirit of collaboration is something that still persists in data journalism today. The first edition of this book was, after all, initially developed by a group of people meeting at the Mozilla Festival in London—and as events around data started to spring up, so did the opportunities for data journalists to work together and share skill sets. If the Iraq and WikiLeaks

releases were great initial examples of cross-Atlantic cooperation, then see how those exercises grew into pan-global reporting involving hundreds of reporters. The Snowden leaks and the Panama Papers were notable for how reporters coordinated around the world to share their stories and build off each other's work.²

Just take an exercise like Electionland, which used collaborative reporting techniques to monitor voting issues in real time on election day. I was involved, too, providing real-time Google data and helping to visualize those concerns in real time. To this date, Electionland is the biggest single-day reporting exercise in history, with over a thousand journalists involved on the day itself. There's a direct line from Electionland to what we were doing in those first few years.

My point is not to list projects but to highlight the broader context of those earlier years, not just at *The Guardian*, but in newsrooms around the world. *The New York Times*, the *Los Angeles Times*, *La Nación* in Argentina: Across the world journalists were discovering new ways to work by telling data-led stories in innovative ways. This was the background to the first edition of this book. *La Nación* in Argentina is a good example of this. A small team of enthused reporters taught themselves how to visualize with Tableau (at that time a new tool) and combined this with freedom of information reports to kickstart a world of data journalism in Latin and South America.

Data journalism went from being the province of a few loners to an established part of many major newsrooms. But one trend became clear even then: Whenever a new technique is introduced in reporting, data would not only be a key part of it but data journalists would be right there in the middle of it. In a period of less than three years, crowdsourcing became an established newsroom tool, and journalists found data, used databases to manage huge document dumps, published data sets and applied data-driven analytical techniques to complex news stories.

This should not be seen as an isolated development within the field of journalism. These were just the effects of huge developments in international transparency beyond the setting up of open data portals. These included campaigns such as those run by Free Our Data, the Open Knowledge Foundation and civic tech groups to increase the pressure on the UK government to open up news data sets for public use and provide APIs for anyone to explore. They also included increased access to powerful free data visualization and cleaning tools, such as OpenRefine, Google Fusion Tables, Many Eyes,

2 For more on large-scale collaborations around the Panama Papers, see Díaz-Struck, Gallego and Romera's chapter in this volume.

Datawrapper, Tableau Public and more. Those free tools combined with access to a lot of free public data facilitated the production of more and more public-facing visualizations and data projects. Newsrooms, such as *The Texas Tribune* and *ProPublica*, started to build operations around this data.

Can you see how this works? A virtuous circle of data, easy processing, data visualization, more data, and so on. The more data is out there, the more work is done with the data the greater pressure there is for more data to be released. When I wrote the piece “Data Journalism Is the New Punk” it was making that point: We were at a place where creativity could really run free (Rogers, 2012). But also where the work would eventually become mainstream.

Data can’t do everything. As Jonathan Gray (2012) wrote: “The current wave of excitement about data, data technologies and all things data-driven might lead one to suspect that this machine-readable, structured stuff is a special case.” It is just one piece of the puzzle of evidence that reporters have to assemble. But as there is more and more data available, that role changes and becomes even more important. The ability to access and analyze huge data sets was the main attraction for my next career move.

In 2013, I got the chance to move to California and join Twitter as its first data editor—and it was clear that data had entered the vocabulary of mainstream publishing, certainly in the United States and Europe. A number of data journalism sites sprouted within weeks of each other, such as *The New York Times’* Upshot and Nate Silver’s *FiveThirtyEight*. Audiences out there in the world were becoming more and more visually literate and appreciative of sophisticated visualizations of complex topics. You will ask what evidence I have that the world is comfortable with data visualizations? I don’t have a lot beyond my experience that producing a visual which garners a big reaction online is harder than it used to be. Where we all used to react with “oohs and aahs” to visuals, now it’s harder to get beyond a shrug.

By the time I joined the Google News Lab to work on data journalism in 2015, it had become clear that the field has access to greater and larger data sets than ever before. Every day, there are billions of searches, a significant proportion of which have never been seen before. And increasingly reporters are taking that data and analyzing it, along with tweets and Facebook likes.³ This is the exhaust of modern life, turned around and given back to us as insights about the way we live today.

Data journalism is now also more widespread than it has ever been. In 2016, the Data Journalism Awards received a record 471 entries. But the 2018

3 For further perspectives on this, see the “Investigating Data, Platforms and Algorithms” section.

awards received nearly 700, over half from small newsrooms, and many from across the world. And those entries are becoming more and more innovative. Artificial intelligence, or machine learning, has become a tool for data journalism, as evidenced by Peter Aldhous' work at *Buzzfeed* (Aldhous, 2017). Meanwhile access to new technologies like virtual and augmented reality open up possibilities for telling stories with data in new ways. As someone whose job is to imagine how data journalism could change—and what we can do to support it—I look at how emerging technologies can be made easier for more reporters to integrate into their work. For example, we recently worked with design studio Datavized to build TwoTone, a visual tool to translate data into sound.⁴

What does a data journalist at Google do? I get to tell stories with a large and rich collection of data sets, as well as getting to work with talented designers to imagine the future of news data visualization and the role of new technologies in journalism. Part of my role is to help explore how new technologies can be matched with the right use cases and circumstances in which they are appropriate and useful. This role also involves exploring how journalists are using data and digital technologies to tell stories in new ways. For example, one recent project, *El Universal's* "Zones of Silence," demonstrated the use of AI in journalism, using language processing to analyze news coverage of drug cartel murders and compare them to the official data, the gap between the two being areas of silence in reporting. I helped them do it, through access to AI APIs and design resources.

The challenges are great, for all of us. We all consume information in increasingly mobile ways, which brings its own challenges. The days of full-screen complex visualizations have crashed against the fact that more than half of us now read the news on our phones or other mobile devices (a third of us read the news on the toilet, according to a Reuters news consumption study (Newman et al., 2017)). That means that increasingly newsroom designers have to design for tiny screens and dwindling attention spans.

We also have a new problem that can stop us learning from the past. Code dies, libraries rot and eventually much of the most ambitious work in journalism just dies. *The Guardian's* MPs' expenses, EveryBlock and other projects have all succumbed to a vanishing institutional memory. This problem of vanishing data journalism is already subject to some innovative approaches (as you can see from Broussard's chapter in this book). In the long run, this requires proper investment and it remains to be seen if the community is sufficiently motivated to make it happen.

4 <https://twotone.io/>

And we face a wider and increasingly alarming issue: Trust. Data analysis has always been subject to interpretation and disagreement, but good data journalism can overcome that. At a time when belief in the news and a shared set of facts are in doubt every day, data journalism can light the way for us, by bringing facts and evidence to light in an accessible way.

So, despite all the change, some things are constant in this field. Data journalism has a long history,⁵ but in 2009, data journalism seemed an important way to get at a common truth, something we could all get behind. Now that need is greater than ever before.

Works Cited

- Aldhous, P. (2017, August 8). We trained a computer to search for hidden spy planes. This is what it found. *BuzzFeed News*. <https://www.buzzfeednews.com/article/peteraldhous/hidden-spy-planes>
- Gray, J. (2012, May 31). What data can and cannot do. *The Guardian*. <https://www.theguardian.com/news/datablog/2012/may/31/data-journalism-focused-critical>
- Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). *Digital News Report 2017*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_o.pdf
- Robertson, C. (2011, December 9). Reading the riots: How the 1967 Detroit riots were investigated. *The Guardian*. <https://www.theguardian.com/uk/video/2011/dec/09/reading-the-riots-detroit-meyer-video>
- Rogers, S. (2012, May 24). Anyone can do it. Data journalism is the new punk. *The Guardian*. <https://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk>

About the Author

Simon Rogers is data editor on the News Lab team at Google, based in San Francisco, he is director of the Sigma Data Journalism Awards and teaches Data Journalism at Medill-Northwestern University.

5 See, for example, the chapters by Anderson and Cohen in this volume.