

Special issue on Data Literacy: Notes from the Field

Don't ask much from Data Literacy

Nicolas Kayser-Bril

Journalism ++, Germany

hi@nkb.fr

It can be argued that data literacy skills are all we need for data literacy to be used properly by journalists. This is wrong. Before skills, journalists need a reason to learn and use data literacy. Such incentives are not in place yet.

“Datajournalism is social science on a deadline”, once said Steven Doig, a Pulitzer-prize winner (Remington, 2012). Indeed, doing journalism with structured data requires that journalists use the social scientist’s toolbox, from data collection to data analysis (Kayser-Bril, 2015). Data literacy is needed at every step of the process: To know what data to collect, understand biases in a data set, perform sensible analyses and visualize the results properly.

Statisticians, once pariah, are now in demand within newsrooms. Many news operations have created dedicated teams to work with structured data. In the United States, the New-York Times is famous for its embrace of datajournalism. Examples abound throughout Europe as well, from public service broadcasters such as Bayerischer Rundfunk (Munich, Germany) or Schweizer Radio und Fernsehen (Zurich, Switzerland) to newspapers of record such as Gazeta Wyborcza (Warsaw, Poland) to local news outlets such as Le Télégramme (Brest, France) or Heilbronner Stimme (Heilbronn, Germany). They published several stories that could never have been made without data literacy skills. Just like muckrakers in the late 19th century forced public institutions to listen to the voiceless, datajournalists today force them to measure the unmeasured (Gray, Lämmerhirt, & Bounegru, 2016).

Data literacy enabled The Guardian to record every person killed by the police in the United States in 2015 with The Counted, for instance. This prompted the FBI to rethink the way it measured the issue. In Spain, Civio systematically analyzed official pardons with the Indultómetro. The number of pardons went from 500 per year to less than 100 a year after the project was published. Journalism++, an agency for datajournalism I co-founded, had similar successes when we measured the gap in what women and men pay (The Woman Tax) or when

Kayser-Bril, N. (2016). Don't ask much from data literacy. *The Journal of Community Informatics*, 12(3), 217–222.

Date submitted: 2016-01-11. Date accepted: 2016-06-13.

Copyright (C), 2016 (the authors as stated). Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5. Available at: www.ci-journal.net/index.php/ciej/article/view/1297

we measured the mortality rates of people trying to come or stay in Europe, with The Migrants' Files (coming to Europe is slightly less dangerous than going into battle at Verdun in 1916). Both projects were instrumental in pushing authorities to measure the issue.

Journalists with data literacy skills are able to produce new (and, by virtue of being unique, exclusive) stories that have an impact and resonate with the audience. It can even be profitable. The Upshot, the New-York Times' datajournalism operation, made up to 5% of the newsroom's page views in 2014 (Sebastian, 2015) with 1.5% of the total staff (Wilson, 2015). Prejudice against math and stats in the newsroom is waning and datajournalism operations are popping up across the world. From this perspective, one could think that data literacy will trickle down the newsroom and, from there, throughout society. But one would be mistaken to do so.

Misusing data

The Migrants' Files, our body counting project, provides a table where each line is a separate incident during which a person died in their attempts to reach or stay in Europe. It aggregates information from a wide variety of sources: coast guards, news reports, social media etc. The table has several columns, such as country in which the event happened, cause of death and source. This table is freely accessible online as a Google Spreadsheet. Anyone with basic data manipulation skills can map, filter and analyze the data in a few minutes.

While the table was reused many times, a few journalists wrongly used the data we collected and came to false conclusions. It happened, for instance, that journalists grouped the data by country to compute a list of most dangerous countries for refugees, even though we made clear, in the data itself and through email contacts, that the data collection process did not allow for such a usage. Our data set has more lines of events that happened in Germany because more organizations there record them. By grouping by country, these journalists wrote that Germany was the country where most refugees died. If anything, Germany might be the safest country for refugees and migrants in Europe.

Using data without data literacy skills is the source of many more errors. It is not rare to see journalists confuse units, for instance between watts, a unit of power, and watt-hours, a unit of energy (Reuters, 2015). It is not rare to read analyses that distort the data to the point where journalists reach conclusions that are precisely opposite to what a data literate person would have said—for example Beale (2015), where an increase in rape reporting is equated to an increase in rape cases.. It is not rare to see articles where billions are used instead of millions. How rarely such mistakes happen is impossible to say. To date, no systematic study of data-driven mistakes has been carried out by academia or professional organizations.

It could be argued that this happens because some people in these newsrooms are still data illiterate. As data literacy increases in the workplace, such mistakes will occur less and less often, the argument goes. It would be true if publishers had an incentive to seek out the truth and root out data illiteracy. They do not.

Wrong incentives

No publisher makes money by publishing facts that are true. Instead, they either sell their readers' attention to advertisers or sell content to their readers, or both. When selling attention, publishers need to garner as much of it as possible. To do so, they do not need to publish true stories, they need to publish articles that will be read and shared while minimizing production costs. This explains why rumor mills abound on the web and why once respectable media outlets like The Daily Mail routinely publish articles that are factually wrong (King, 2015). The lack of data literacy of the staff does not help, but managers have no need for data literate writers. And writers with data literacy skills have no incentive to use them.

Selling content directly to readers does not fundamentally change a publisher's incentives. Instead of volume, a publisher needs to please the audience as much as possible, even when facts need to be distorted to fit the audience's beliefs. Fox News, a cable TV station in the United States, is famous for reporting and creating lies, often with help of wrong charts and statistics (Shere & Groch-Begley, 2012). The news operation derives most of its income not from advertising, but from subscriptions (Shere & Groch-Begley, 2012). It grew by accompanying the rise of reactionary Republicans in the United States and feeds on their disregard for facts.

Still, in the absence of systematic measurement of mistakes done with data, one could argue that a general lack of data literacy skills are more responsible for bad reporting of numerical information than misplaced incentives. The argument does not hold when one looks at other fields where data literacy is widespread and incentives as misaligned.

Testimonies from financiers, many of whom have degrees in statistics from top-tier universities, show how wrong incentives can make one knowingly ignore their data literacy skills. An employee of Standard & Poors, a company that rates the risk of financial products, once said that they would rate any securities, even if they were "structured by cows". The reason is given by an employee of Moody's, another ratings agency, when he explained that the errors committed in the run up to the financial crisis "made [them] look either incompetent at credit analysis or like [they] sold [their] soul to the devil for revenue". That high-level staffers of the world's top finance institutions be incompetent is unlikely. The explanation why data literate people put blinders on and pretended not to understand the data before their eyes is given by another Standard & Poor's employee, who said before the financial crisis: "Let's hope we are all wealthy and retired by the time this house of cards falters". All quotes from this paragraph come from the House of Representatives - Hearing Committee on Oversight and Government Reform (2008).

In finance as in journalism, greed trumps data literacy. In academia, too, stories abound where renowned professors manipulate data to increase their standing, gain public visibility (the Seralini affair is a good example) or make money (Goldenberg, 2015; Proctor, 2012). A meta study of data falsification found that a significant chunk (about 10%) of academics had manipulated data at some point (Fanelli, 2009).

These examples show that, while data literacy skills are needed to correctly work with data, they can easily be turned off by other incentives to the point where part (academia) or most

(finance) actions purporting to be driven by data are actually driven by entirely different factors. It would be foolish to believe that journalists are different and that, confronted with conflicting incentives, they will resist and keep looking for the truth in data.

The good data literacy can do

Current incentives do encourage journalists to put data literacy skills to use when their audience is interested in facts. Newsrooms that set up datajournalism teams are mostly the newspaper of record of a given market. Tabloids and other mass-market media outlets have yet to invest in data literacy. They do use data and visualizations to serve “truthiness”, to give a veneer of seriousness to lies, as Fox News exemplifies. An experiment conducted by Italian researchers on a large sample (close to 10,000 Facebook users) showed that factually wrong news items were in great demand by groups exhibiting distrust of established institutions (Bessi, et al., 2015). As distrust in institutions increases in many places (think of Trump’s or Le Pen’s popularity), the market for factually inaccurate news grows. In an article announcing the end of a fact-checking column, The Washington Post’s Caitlin Dewey (2015) rightly asks “Is it the point at which we start segmenting off into alternate realities?” Only one of these realities, which might not be the largest, is interested in data literacy.

Data literacy does enable some journalists to make better, more efficient reporting. Anyone interested in quality journalism should undoubtedly support this trend. But we should not fool ourselves and believe that this will result in better information overall. It will, at most, impact the minority of the public that is interested in developing a fact-based world view.

Teaching data literacy to the general population, by changing the primary school curriculum, for instance, would make little difference. The desire for a person to apprehend the world through facts and consume content from data literate journalists is probably not function of the curriculum. If it were, the development of curricula focused more on concrete skills throughout the 20th century (think basic science versus Latin and ancient Greek or versus an absence of education) would have created populations more willing to embrace a fact-based worldview. It did not. Evidence for the disconnect between facts and majority opinion can be found most easily on immigration topics, where the disconnect between facts and public discourse is strongest. Work by Hein de Haas (2014), among many others, exemplifies this – for example.

To achieve the larger goal of engaging citizens with science and data, journalists need more than data literacy skills. They need a reason to acquire and use them. In *Code and Other Laws of Cyberspace*, Lawrence Lessig (1999) writes that socially desirable behaviors can be influenced by markets, law, code or norms. Many consumers are uninterested in data literate journalists. The Daily Mail and Fox News’ successes are proof of this. Markets will not work. Legislation would not either. Prohibiting a wrongful usage of data would necessarily infringe on freedom of speech, making the cure worst than the illness. Some tried to use code to address the issue. Trooclick, a French startup, created a browser plug-in that automatically looked for mistakes in news items (it didn’t work) (Wilner, 2015). Beyond the enormous technical challenges, the lack of public interest in the issue will not entice computer scientists to automate data literacy. The last option, in Lessig’s framework, are norms. Journalists could be incentivized to become data literate through encouragement (by setting up more datajournalism prizes, for instance) and through deterrents, such as the condemnation of data

illiteracy by civil society groups. Scientists could gather and create a media watchdog dedicated to data literacy, for instance. It would be a colossal and expensive endeavor, but, absent a change in the context that fosters data illiteracy (distrust in institutions), this is the only available option.

References

- Beale, C. (2015). *Three times as many rapes in Delhi since 2012 gang rape*. Retrieved from Daily Telegraph: <http://www.telegraph.co.uk/news/worldnews/asia/india/12052095/Three-times-as-many-rapes-in-Delhi-since-2012-gang-rape.html>
- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs Conspiracy: Collective Narratives in the Age of Misinformation. *PLoS ONE*. Retrieved from <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118093>
- de Haas, H. (2014). *Human Migration: Myths, Hysteria and Facts*. Retrieved from <http://heindehaas.blogspot.co.uk/2014/07/human-migration-myths-hysteria-and-facts.html>
- Dewey, C. (2015). *What was fake on the Internet this week: Why this is the final column*. Retrieved from The Washington Post: <https://www.washingtonpost.com/news/the-intersect/wp/2015/12/18/what-was-fake-on-the-internet-this-week-why-this-is-the-final-column/>
- Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE*. Retrieved from <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738>
- Goldenberg, S. (2015). *Greenpeace exposes sceptics hired to cast doubt on climate science*. Retrieved from The Guardian: <https://www.theguardian.com/environment/2015/dec/08/greenpeace-exposes-sceptics-cast-doubt-climate-science>
- Gray, J., Lämmerhirt, D., & Bounegru, L. (2016). Changing What Counts: How Can Citizen-Generated and Civil Society Data Be Used as an Advocacy Tool to Change Official Data Collection? *SSRN*.
- House of Representatives - Hearing Committee on Oversight and Government Reform. (2008). *Credit Rating Agencies and the Financial Crisis*. Retrieved from <https://www.gpo.gov/fdsys/pkg/CHRG-110hhrg51103/html/CHRG-110hhrg51103.htm>
- Kayser-Bril, N. (2015, September). *Data Journalism*. Retrieved from <http://blog.nkb.fr/datajournalism>
- King, J. (2015). *My Year Ripping Off the Web With the Daily Mail Online*. Retrieved from <http://tktk.gawker.com/my-year-ripping-off-the-web-with-the-daily-mail-online-1689453286>
- Lessig, L. (1999). *Code and Other Laws of Cyberspace*. New York: Basic.
- Pew Research Center. (2014). *Cable TV: Revenue Streams by Channel*. Retrieved from <https://web.archive.org/web/20160422034905/http://www.journalism.org/media-indicators/revenue-streams-by-cable-news-channel/>
- Proctor, R. (2012). *Golden Holocaust. Origins of the Cigarette Catastrophe and the Case for Abolition*. Berkeley and Los Angeles, California: University of California Press.
- Remington, A. (2012). Retrieved from Journalist's Resource: <http://journalistsresource.org/tip-sheets/research/research-chat-steve-doig-data-journalism-social-science-deadline>
- Reuters. (2015). *EUROPE POWER-Abundant wind power curbs German spot prices in low demand week*. Retrieved from <http://af.reuters.com/article/commoditiesNews/idAFL8N14A2XD20151221>
- Sebastian, M. (2015). *The Upshot Emerges as Potentially Lucrative Franchise at The Times*. Retrieved from AdAge: <http://adage.com/article/media/upshot-a-potentially-lucrative-franchise-times/296616/>

- Shere, D., & Groch-Begley, H. (2012). *A History Of Dishonest Fox Charts*. Retrieved from Media Matters for America: <http://mediamatters.org/research/2012/10/01/a-history-of-dishonest-fox-charts/190225>
- Wilner, T. (2015). *'There is No Market for Fact-Checking': Trooclick Exits the Verification Scene*. Retrieved from <https://medium.com/@tamarwilner/there-is-no-market-for-fact-checking-trooclick-exits-the-verification-scene-4488565bf06c#.nbpbd675i>
- Wilson, M. (2015). *The Upshot: Where The New York Times Is Redesigning News*. Retrieved from Fast Company: <https://www.fastcodesign.com/3040817/the-upshot-where-the-new-york-times-is-redesigning-news>